Journal of Marketing Analytics
https://doi.org/10.1057/s41270-023-00274-y

ORIGINAL ARTICLE

Time analysis of online consumer behavior by decision trees, GUHA association rules, and formal concept analysis

Tomáš Pitka[1] · Jozef Bucko[1] · Stanislav Krajči[2] · Ondrej Krídlo[2] · Ján Guniš[2] · Ľubomír Šnajder[2] · Ľubomír Antoni[2] · Peter Eliaš[3]

Revised: 4 July 2023 / Accepted: 20 November 2023

Abstract
Data analytics plays a significant role within the context of the digital business landscape, particularly concerning online sales, aiming to enhance understanding of customer behaviors in the online realm. We review the recent perspectives and empirical findings from several years of scholarly investigation. Furthermore, we propose combining computational methods to scrutinize online customer behavior. We apply the decision tree construction, GUHA (General Unary Hypotheses Automaton) association rules, and Formal concept analysis for the input dataset of 9123 orders (transactions) of sports nutrition, healthy foods, fitness clothing, and accessories. Data from 2014 to 2021, covering eight years, are employed. We present the empirical discoveries, engage in a critical discourse concerning these findings, and delineate the constraints inherent in the research process. The decision tree for classification of the year's fourth quarter implies that the most important attributes are country, gross profit category, and delivery. The classification of the morning time implies that the most important attributes are gender and country. Thus, the potential marketing strategies can include heterogeneous conditions for men and women based on these findings. Analyzing the identified groups of customers by concept lattices and GUHA association rules can be valuable for targeted marketing, personalized recommendations, or understanding customer preferences.

Keywords Formal concept analysis · Decision trees · GUHA association rules · Consumers · Marketing

Introduction

Analyzing consumers' online purchasing habits can provide many advantages for commercial entities, marketing professionals, and consumers. Classifiers based on decision trees can be applied to predict market trends, specifically for determining when to buy or sell. Decision trees are the transparent and efficient option for machine learning because they sort data attributes at each node to arrive at a decision (Samarth 2023; Vaca et al. 2020). Several researchers proposed a system to enhance the shopping experience at fashion retail outlets. This system smartly groups together various items and customer profiles, which it encounters online and in physical stores. It leverages the power of mining association rules to foresee the shopping patterns of new customers (Bellini et al. 2023; Fan et al. 2023).

✉ Tomáš Pitka
tomas.pitka@tuke.sk

Jozef Bucko
jozef.bucko@tuke.sk

Stanislav Krajči
stanislav.krajci@upjs.sk

Ondrej Krídlo
ondrej.kridlo@upjs.sk

Ján Guniš
jan.gunis@upjs.sk

Ľubomír Šnajder
lubomir.snajder@upjs.sk

Ľubomír Antoni
lubomir.antoni@upjs.sk

Peter Eliaš
elias@saske.sk

[1] Department of Applied Mathematics and Business Informatics, Faculty of Economics, Technical University of Košice, B. Němcovej 32, 04001 Košice, Slovakia

[2] Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 04001 Košice, Slovakia

[3] Mathematical Institute, Slovak Academy of Sciences, Grešákova 6, 04001 Košice, Slovakia

Published online: 09 January 2024

Formal concept analysis (FCA) (Carpineto and Romano 2004; Ganter and Wille 1999; Poelmans et al. 2013) is a method of data analysis based on a lattice theory that has great potential to study how people behave online. This method offers visualization capabilities to explore the relationships in the object-attribute tables that are present, for example, by customer transactions and their characteristics. Combining well-known methods like decision trees and pattern-finding has made these studies more attractive to a broader audience.

The main objective of this paper is to show what is currently known about this topic and how data analysis tools are used to understand customer behavior and buying patterns online. In this paper, we explore online customer behavior by combining decision trees, GUHA (General Unary Hypotheses Automaton) association rules, and Formal concept analysis. We present the findings from these different methods and the practice recommendations. The paper is organized as follows. In "Literature review" section, we review recent studies on analyzing customer behavior with the mentioned data analysis methods. In "Methodology and dataset description" section, we describe data, its source quality, and used methods, including decision trees, GUHA association rules, and concept lattices from FCA. Our findings are presented in the "Results" section. The paper ends with a discussion, recommendations, and notes on its limitations.

Literature review

This section reviews studies on analyzing consumer behavior using decision trees, association rules, and FCA.

Related studies on decision trees

Decision trees are versatile predictive models widely used in artificial intelligence and machine learning to extract hidden patterns from data. They use a logical, rule-based approach to represent and classify events, aiding prediction. Decision trees are particularly effective for categorizing large data sets (Charbuty and Abdulazeez 2021). These studies detail the fundamentals of decision trees and their applications and evaluate specific algorithms, data sets, and results. Yawata et al. (2022) improved decision tree accuracy by adding multi-dimensional boundaries. However, its performance on larger, more complex data sets remains untested, and it has yet to be compared with other advanced methods (Yawata et al. 2022).

Decision trees can also be helpful in remote sensing for distinguishing crop types and patterns, demonstrating their versatility with varied data sets (Tariq et al. 2022). Wang et al. (2022) empirically tested a model linking online trust with consumer e-commerce engagement, suggesting that trust and perceived value significantly influence usage intentions. The study also found that women are more likely than men to use e-commerce for online shopping, providing insights for businesses to enhance their e-commerce strategies (Wang et al. 2022). Charandabi and Ghanadiof (2022) highlight trust's critical role in e-commerce, addressing security and privacy concerns. Despite its importance, defining trust and its effects remains challenging. Their study introduces a Fuzzy Data Envelopment Analysis model to evaluate trust and resilience in online and offline shopping, offering a new perspective on customer behavior in digital markets and suggesting ways to increase customer satisfaction and loyalty (Charandabi and Ghanadiof 2022).

Lastly, decision trees can enhance customer satisfaction by identifying key purchasing factors, allowing businesses to tailor the customer experience and potentially increase sales and loyalty (Samarth 2023). Ganar and Hosein (2022) crafted a composite model to forecast whether a bank's customer is likely to switch to using online banking services. The authors compare two supervised learning algorithms: Decision Tree and Extreme Gradient Boosted (XGBoost). The purpose is to predict clients' behavior when transitioning to an online banking platform. The results indicated that combining the K-Modes clustering algorithm and the XGBoost classification model yielded the best test accuracy, a remarkable 96.1% (Ganar and Hosein 2022).

Wen (2023) analyzed customer churn in banks using logistic regression and decision tree models, with the latter proving more precise and unbiased. Key predictors of churn included age, salary, and product usage, with certain customer profiles showing a higher tendency to leave. They recommend that banks refine their systems to retain customers at high risk of churning (Wen 2023). Similarly, Luo (2023) study employed a Decision Tree Model to forecast churn in telecom firms, finding that customers with long-term contracts are less prone to churn. Factors such as age, marital status, and financial reliance were linked to increased churn. Luo's model also offered specific strategies to reduce customer turnover (Luo 2023).

Tundo and Mahardika (2023) study introduces a method for forecasting roof tile production in Kebumen by combining a Tsukamoto Fuzzy Inference System with a Decision Tree C 4.5. This research tackles the issue of production management for tile manufacturers, aiming to align output with customer demand and optimize profits. Traditional systems often result in either excess or insufficient production. The authors propose a computerized forecasting tool that employs Tsukamoto fuzzy logic, which generates rules from the C 4.5 decision tree without expert input. The C 4.5 algorithm forms these rules by learning from data-capturing conditions that frequently arise. The study's findings indicate that the model's forecasts were close to the actual production numbers, with a 29.34% error rate and 70.66% accuracy, as per the Average Forecasting Error Rate. The model proved to be very effective, fulfilling all customer orders. Its predictions were used to determine production volumes or were integrated with inventory levels to meet customer needs fully (Tundo and Mahardika 2023).

Related studies on association rules

Analyzing customer behavior using association rules involves data mining techniques to uncover complex patterns in customer purchases. These patterns provide insights into how different products or services are related from the customer's point of view. Such information is crucial for companies to fine-tune marketing strategies, decide on product placement, and improve inventory management. Fan et al. (2023) presented a new method to identify association rules for unsafe behavior in various industrial settings using comprehensive data. They applied the Apriori algorithm to explore patterns of hazardous behavior in subway construction, finding significant rules that link the type of work, construction stage, work hours, and unsafe practices (Fan et al. 2023). The authors could have expanded their research to understand why certain stages or times are prone to unsafe actions. Comparing their method with others in representing and analyzing unsafe behavior data could have validated their approach further.

The following study by Husein et al. (2022) discussed the challenges of predicting customer segments in e-commerce using traditional models. They proposed a clustering method to categorize customer data, using the Apriori algorithm to identify common purchasing patterns. Their work demonstrates the effectiveness of the K-Means Clustering technique combined with the RFM model in recognizing patterns of product purchases (Husein et al. 2022). The authors clearly outline the problem, their solution, and the evaluation of their method, contributing to the current understanding of customer segmentation.

The study by Zhou et al. (2020) delves into the enhancement of warehouse order-picking efficiency through the application of big data. The research involved the collection and analysis of data to improve storage organization through the implementation of clustering and association algorithms for item categorization. The refined approach to the conventional ABC storage method resulted in improved accuracy in goods classification. The study concludes that the application of intelligent data analysis can optimize warehouse operations, leading to cost savings without the need for additional storage space. Thus, the study provides valuable insights into the potential of big data to enhance warehouse management and make it more efficient (Zhou et al. 2020).

Xiao and Piao (2022) explored the significant data era's onset and how mobile edge computing's transformative potential has led to a surge in consumer data production. They highlighted the critical need to analyze this vast data for valuable insights. The study focused on understanding customer groups' consumption patterns to develop effective marketing strategies, particularly in the telecom industry. It pointed out the growing use of association rules for analyzing customer behavior due to the complex data and algorithms involved. The paper noted that while association rules and mobile computing are potent for identifying patterns in large data sets, they have limits, such as the inability to infer causality (Xiao and Piao 2022).

Stuti et al. (2022) presented a detailed method to extract insights from customer purchasing data to improve product recommendations. Their two-step approach identified product correlations from user transactions and then formulated utility-based association rules to map purchasing trends. These rules were integrated into a recommender system to suggest new products. The study also evaluated the accuracy of these rules against those derived from Frequent Item Set Mining and Improved Utility-Based Mining on an e-commerce platform (Stuti et al. 2022).

Roy (2016) explored the retail industry in India, mainly focusing on the food and grocery segment. The study aimed to assess the association between shopping trip patterns and consumer basket size and identify factors influencing a customer's choice of store format for purchasing grocery items. The study discusses how internal factors like emotions and external factors like store location influence consumer behavior. Demographic factors, convenience, and merchandise assortment influence grocery store choice behavior. The study found that respondents who purchase items less frequently (fortnightly or monthly) tend to shop at hypermarkets or departmental stores to take advantage of offers and discounts. In contrast, those who shop daily prefer convenience or Kirana stores for the credit facilities and personal relationships they offer. The article concludes that both the purchase pattern and the volume significantly influence the choice of store format for grocery shopping in Kolkata (Roy 2016).

Related studies on FCA

Formal concept analysis is a mathematical approach used for data analysis and knowledge representation, widely utilized in fields like customer segmentation. It is a computational method grounded in the theory of partial orders, which excels at uncovering the structural properties of binary data. FCA is particularly adept at identifying how different subsets within a dataset are interrelated. The method is highly regarded for its applications in various areas, especially bioinformatics, which is crucial for analyzing and making sense of complex data patterns.

The research by Roscoe et al. (2022) highlights FCA's role in revealing the complex interplay and dependencies found in gene expression data, thereby contributing significantly to bioinformatics. In gene expression, which sheds light on how genes are activated under different conditions, FCA is an essential tool for a methodical analysis of such data, aiding in disease prediction and understanding vital biological processes (Roscoe et al. 2022).

Customer segmentation involves categorizing a company's customers into groups based on common characteristics like age, gender, interests, and purchasing patterns. Using the FCA method to investigate customer behavior and their classification represents a new approach to this issue that offers fresh perspectives and methodologies. Rungruang et al. (2024) introduce a novel customer segmentation technique that blends the RFM model with FCA. Their method bridges the gap between complex data science techniques and their practical business applications. This approach segments customers and reveals the underlying patterns in customer data, offering insights that traditional clustering methods do not. They tested their model against standard clustering techniques using a well-known dataset and found it provides marketers with clear, actionable insights for crafting marketing strategies. The paper delves into the RFM model, which categorizes customers based on purchase recency, frequency, and spending, and describes how to score customers using this model. In summary, the research offers a new clustering algorithm that leverages RFM and FCA for practical, real-world business segmentation (Rungruang et al. 2024).

Kwon et al. (2023) delve into mobile app usage patterns, examining them through the lens of media combination. They unveil a new technique called Weighted Formal Concept Analysis (WFCA) to discern standard sets of apps, known as repertoires and standout apps, or 'killer apps,' focusing on demographic differences. Their findings highlight that communication and lifestyle apps are universally popular among all demographics. The study employs FCA to map out these app repertoires and to explore the concept of mobile app repertoires more deeply. The authors point out that FCA offers a systematic way to decode the intricate web of app usage patterns across various demographic groups. By integrating weights into FCA, the WFCA method enhances this exploration, providing a detailed picture of user preferences and behaviors (Kwon et al. 2023).

Yang et al. (2019) developed new, eco-friendly financial products for China's rapidly growing economy. They used a particular method called a fuzzy concept lattice, which organizes data without duplication and emphasizes essential features. This makes the method less complex and improves the evaluation of data. They adapted this method to work better with different data types, created a new algorithm, and showed how it works with green finance data. Their research led to a practical blueprint for creating green financial products (Yang et al. 2019).

Methodology and dataset description

Before beginning any analysis, describing the input data thoroughly is crucial. This initial step provides essential context and insight into the data under scrutiny, enabling the analyst to spot any potential issues or biases and to confirm that the data are appropriate for the chosen analytical methods (Schröer et al. 2021). Moreover, detailing the input data helps peers who may examine the analysis to grasp the origins and characteristics of the data utilized. A clear description of the input data is vital for maintaining the analysis's validity and precision. The CRISP-DM framework, widely recognized in the industry for data mining, is often used to analyze and interpret consumer trends (Plotnikova et al. 2021).

Preparing data is a crucial and often the lengthiest phase in data mining, essential for obtaining pertinent and insightful outcomes. This preparation involves cleansing, labeling, and structuring unprocessed data, rendering it apt for machine learning and analysis. It also encompasses exploring and charting the data to unearth patterns. Remarkably, data preparation might consume up to 80% of project time, underscoring the need for specialized tools to streamline this phase (Saltz and Krasteva 2022). In our scenario, data preparation commenced with acquiring the data from a company, which would serve as a primary research site for our design applications. Given the extensive dataset with numerous parameters, grasping the data's essence was a pivotal initial step in our dissertation. The initial data encompassed databases detailing about

- orders from the pertinent reporting period,
- customer profiles,
- product details,
- goods procurement information.

At the outset, we focused on comprehending the data's variables and structure. We can categorize the data processing journey into distinct stages:

- grasping and initial processing of data,
- verification and validation of data,
- visualization of data.

In data mining, the CRISP-DM (Cross-Industry Standard Process for Data Mining) is a leading framework used to uncover customer behavior patterns. It includes the following phases (Schröer et al. 2021):

Business understanding During this phase, we aimed to grasp the business objectives and requirements of the project. We engaged in frequent virtual meetings with relevant departments, using accurate organizational data to guide the project's direction and outcomes. However, this critical step must be addressed in the CRISP-DM process, leading to vague goals or repetitive results. Acknowledging the significance of this phase is essential for successful research.

Data understanding A crucial step in the CRISP-DM process is understanding the data (Plotnikova et al. 2021). This phase is closely linked to the first stage, as the data's nature often sets the research's scope and possible findings. Our work involved getting to know the datasets, which included detailed customer, order, and product information from 10 different databases. Understanding the relationships and dynamics between these databases was vital for progressing our research. Like the first phase, our in-depth discussions with company department representatives focused on these databases. A solid understanding of the data is necessary to explore customer behavior thoroughly, and skipping this step could undermine the integrity and accuracy of our results. Central to understanding the data was identifying the accuracy of the linkage identifier across each database. Before meeting with company officials, we tried to link the databases using key identifiers like order ID, customer ID, email, entity ID, and product ID. We aimed to create a unified database with all relevant variables for our study. Initially, we used the entity ID as a familiar identifier, but discussions with company stakeholders revealed that the entity ID varied across databases, making it unreliable for our purposes. Recognizing the importance of accurate data understanding, we combined the databases using multiple identifiers to ensure precise variable alignment in the central database for the following CRISP-DM stages.

Data preparation The initial stage focused on primary data operations such as cleaning, reducing, transforming, integrating, and creating new variables. After thoroughly understanding the data, we reviewed it analytically using descriptive statistics to explore variable distributions, attribute descriptions, and the overall data structure. This detailed examination helped us move smoothly to the next CRISP-DM phase: modeling. With these insights, we improved our data preparation methods, which enhanced the quality of our modeling results. Additionally, we created new variables like time of day, day of the week, fiscal quarter, and delivery method during this data processing phase.

Modeling During this stage, we utilized various models to uncover patterns and connections in consumer behavior. This involved choosing the most appropriate model and then assessing its performance.

Data description and source quality

GymBeam, the fitness e-commerce platform with nutritional supplements, provides and maintains our dataset, empowering academic researchers to explore crucial aspects of e-shop consumer behavior in Eastern Europe. The anonymized dataset contains 9123 orders (transactions) of sports nutrition, healthy foods, fitness clothing, and accessories. Data from 2014 to 2021, covering 8 years, are employed.

The set of 19 categorical attributes is available in our dataset with no missing values. In particular, order_ID attribute is a primary key in our data and uniquely identifies each order. The dataset contains 70 product categories (attribute category_ID), ranging from 1 to 1,911 orders for each category value and an average of 130.3. The gross profit category includes three possible classes: A (56.2%), B (23.7%), and C (20.1%). We consider delivery and pickup for the categories of shipping service attribute. The time of order, e.g., day name, hour, week, and other time attributes, is included in the dataset.

Attribute	No. of categories	Attribute	No. of categories
order_ID	9123	day_of_month	31
gross_profit_category	3	month	12
gender	2	year	8
country	8	hour	24
shipping_service	2	day_name	7
bestseller	2	day_part	6
long_term_unavailable	2	quarter_of_year	4
product_ID	70	week	53
Q4_period	2	day_of_year	365
		morning	2

Looking at the relevance of the input data, the company takes great care in how its data are collected. As mentioned several times in the text, the input database contains parameters related to the company's orders and customers. These raw data were processed during the CRISP-DM process, specifically in the data preparation part. For example,

Unification of order status Orders progress through various stages: pending, prepared, shipped, in transit, delivered, and returned, among others. The order is deemed complete when its final status is "delivered." Focusing only on orders marked as "delivered" and excluding every other type of order status allows us to identify transactions that reached customers accurately. This approach helps to remove any bias from our analysis that could arise from including orders that were never received or were returned.

Consolidation of the gender attribute The company uses two approaches to determine the customer's gender. The first represents a situation where a registered customer places the order. In this case, the customer fills in the gender parameter himself. When a non-registered user makes an order, the company uses an algorithm to predict the customer's gender with a prediction accuracy between 80% and 95%. However, this value was not included for a specific group of orders, or the % accuracy needed to be higher. Such orders were excluded from the final database due to concerns of data bias.

Based on the information on the origin of the data received from the company itself and the procedures we have carried out related to CRISP-DM, we can consider our input data as relevant and ready to enter the modeling phase.

In the following part of this section, we present the selected basic notions of three methods (decision trees, GUHA method, and formal concept analysis), which we applied to our dataset.

Decision trees

In this subsection, we will present the classification method of decision trees, which we apply for the classification of orders by quartiles or day parts. In classification tasks, the values of input categorical attributes are called categories (or classes). The values of the target attribute are usually called labels.

The classification method of decision trees is based on the recursive division of the training set into two or more parts (Quinlan 1992; Breiman 1996). A mathematical structure of a tree is created by selecting one input attribute for splitting in each iteration. The splitting attribute is selected using entropy, Gini index, correlation coefficients, statistical tests, and other criteria. The aim is to increase the homogeneity of the subsets of objects concerning the values of the target attribute. Each leaf node of the tree (corresponding to a concatenation of logical conditions on the input attributes) is labeled by the most frequent category of the target attribute in the subset of objects. The tree's construction is completed if the objects of the same target attribute category (class) are present in the tree's leaves or if the stopping criterion is fulfilled (e.g., the minimal number of objects in the leaf node or the depth of the tree).

From a mathematical point of view, we can consider a set A of input categorical attributes and $b \notin A$ of a target categorical attribute. As an illustrative example from our data, we can take the set of input attributes

$$A = \{\text{gender, shipping\_service, bestseller}\},$$

and a target attribute

$$b = \text{Q4\_period}.$$

A set L_a for each $a \in A$ expresses the categories of the input attribute, and a set L_b for b represents the categories of the target attribute. In our example, we have that

$$\begin{aligned} L_{\text{gender}} &= \{\text{woman,man}\}, \\ L_{\text{shipping\_service}} &= \{\text{delivery,pickup}\}, \\ L_{\text{bestseller}} &= \{\text{yes,no}\}, \\ L_{\text{Q4\_period}} &= \{\text{Q4\_yes,Q4\_no}\}. \end{aligned}$$

In general, the **input** for decision tree construction can be expressed as a training set D such that it holds

$$D \subseteq L_{a_1} \times \ldots \times L_{a_n} \times L_b$$

for a set $A = \{a_1, \ldots, a_n\}$ of input attributes and a target attribute b.

Given a training set, we describe a general method for constructing a decision tree in Algorithm 1.

Algorithm 1 Decision_tree_construction(D, minProportion)

```
Require: A training set $D$, minProportion $\in (0.5, 1]$
Ensure: A decision tree
if minProportion of objects in $D$ are all of the same class of $L_b$ (i.e., termination condition) then
    Create a leaf node $v$ labeled with the majority class in $D$
else
    for each attribute $a \in A$ do
        Compute the quality of $a$ with respect to the target attribute $b$ by splitting criterion (e.g., entropy or Gini index)
    end for
    $a^* \Leftarrow$ The splitting attribute with the best quality
    $A \Leftarrow A \setminus \{a^*\}$
    for $j \in L_{a^*}$ do
        $D_j \Leftarrow$ the set of all objects from $D$ with category $j$
        Decision_tree_construction($D_j$, minProportion)
    end for
end if
```

We will explain Algorithm 1 in more detail to help readers understand the application in our research. In the following example, consider the training set D and value 0.8 of minProportion input parameter in Algorithm 1.

D	Gender	Shipping_service	Bestseller	Q4_period
Order #36	Man	Pickup	No	Q4_no
Order #47	Woman	Delivery	No	Q4_no
Order #128	Man	Delivery	Yes	Q4_no
Order #151	Woman	Delivery	Yes	Q4_yes
Order #160	Man	Delivery	Yes	Q4_yes
Order #199	Woman	Pickup	Yes	Q4_yes
Order #401	Man	Pickup	Yes	Q4_yes

We have seven orders (transactions) in our example of the training set D. In the first line of Algorithm 1, we need to check the proportion of orders in Q4 period. Let $y \in L_b$, D be the training set, and D_y be the set of all objects from D with category y. The mapping

$$t_y = \frac{|\mathrm{D}_y|}{|\mathrm{D}|} = \frac{|\{(x_1, \dots, x_n, y) \in D : (x_1, \dots, x_n) \in L_{a_1} \times \dots \times L_{a_n}\}|}{|\mathrm{D}|}$$

corresponds to the proportion of objects from D with target attribute category y. In our example, we have that

$$|\mathrm{D}| = 7,$$
$$|\mathrm{D}_{\text{Q4\_yes}}| = |\{\text{order\#151}, \text{order\#160}, \text{order\#199}, \text{order\#401}\}| = 4,$$
$$|\mathrm{D}_{\text{Q4\_no}}| = |\{\text{order\#36}, \text{order\#47}, \text{order\#128}\}| = 3.$$

Thus, we have

$$t_{\text{Q4\_yes}} = \frac{4}{7} = 0.57, \qquad t_{\text{Q4\_no}} = \frac{3}{7} = 0.43.$$

Since 57% of orders are in Q4 period and 43% are not, the termination condition is not fulfilled (minProportion = $0.8 > 0.57$), and we continue with line 4 of Algorithm 1.

In the fourth line of Algorithm 1, we compute the quality of each input attribute by splitting criterion (e.g., entropy or Gini index). These measures describe the homogeneity in the subset of objects regarding the target attribute. A smaller entropy or Gini index value corresponds to higher homogeneity in the given subset of objects.

First, we will formally describe the evaluation of the quality of input attributes by entropy or Gini index for a training set D. For a selected input attribute $a \in A$ and its category $z \in L_a$, we can define a subset $D^{a,z}$ of a training set D such that it includes all objects of category z of attribute a. For example,

$$D^{\text{bestseller,yes}} = \{\text{order\#128}, \text{order\#151}, \text{order\#160}, \text{order\#199}, \text{order\#401}\}$$

and

$$D^{\text{bestseller,no}} = \{\text{order\#36}, \text{order\#47}\}.$$

Analogously, we can define a mapping $t^{a,z}$ for the proportion of objects from D with the category z of the input attribute a.

Let $y_i \in L_b$ for all $i \in \{1, \dots, |L_b|\}$. A tuple

$$\mathcal{T}^{D^{a,z}} = \left(t^{a,z}_{y_1}, \dots, t^{a,z}_{y_{|L_b|}}\right)$$

corresponds to a discrete probability distribution for a subset $D_a^{a,z}$ of a training set D, whereby

$$\sum_{y \in L_b} t^{a,z}_y = 1.$$

In our example, we have that $L_b = \{\text{Q4\_yes}, \text{Q4\_no}\}$ and

$$t^{\text{bestseller,yes}}_{\text{Q4\_yes}} = \frac{4}{5} \qquad t^{\text{bestseller,yes}}_{\text{Q4\_no}} = \frac{1}{5}$$

Thus,

$$\mathcal{T}^{D^{\text{bestseller,yes}}} = \left(\frac{4}{5}, \frac{1}{5}\right)$$

.

Now, for a given $\mathcal{T}^{D^{a,z}}$, an **entropy** is defined by a function

$$\text{entropy}(\mathcal{T}^{D^{a,z}}) = -\sum_{y \in L_b : t^{a,z}_y \neq 0} t^{a,z}_y \log_2 t^{a,z}_y.$$

and Gini index is defined by a function

$$\text{gini}(\mathcal{T}^{D^{a,z}}) = 1 - \sum_{y \in L_b} (t^{a,z})^2.$$

We will explain the computation of entropy and the overall quality of attribute bestseller in our running example. We have that

$$\text{entropy}\left(\mathcal{T}^{D^{\text{bestseller,yes}}}\right) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722,$$
$$\text{entropy}\left(\mathcal{T}^{D^{\text{bestseller,no}}}\right) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0.$$

To evaluate the overall quality of the bestseller attribute, the weighted average of the previous two values of entropy will be computed by

$$\text{entropy\_average(bestseller)} = \frac{5}{7} \cdot 0.722 + \frac{2}{7} \cdot 0 = 0.516$$

since we have four bestsellers and two not bestsellers out of six orders.

Analogously, we can evaluate other input attributes in our running example. Thus, we have that

$$\text{entropy\_average(gender)} = \frac{4}{7} \cdot 1 + \frac{3}{7} \cdot 0.918 = 0.965,$$
$$\text{entropy\_average(shipping\_service)} = \frac{4}{7} \cdot 1 + \frac{3}{7} \cdot 0.918 = 0.965.$$

The bestseller attribute obtained the smallest value of the weighted average of entropies (i.e., the highest homogeneity in the subset of objects). Thus, we have $a^* =$ bestseller in lines 7 and 8 of Algorithm 1. Moreover, in Lines 9–11, we will split the root of our decision tree into two nodes by

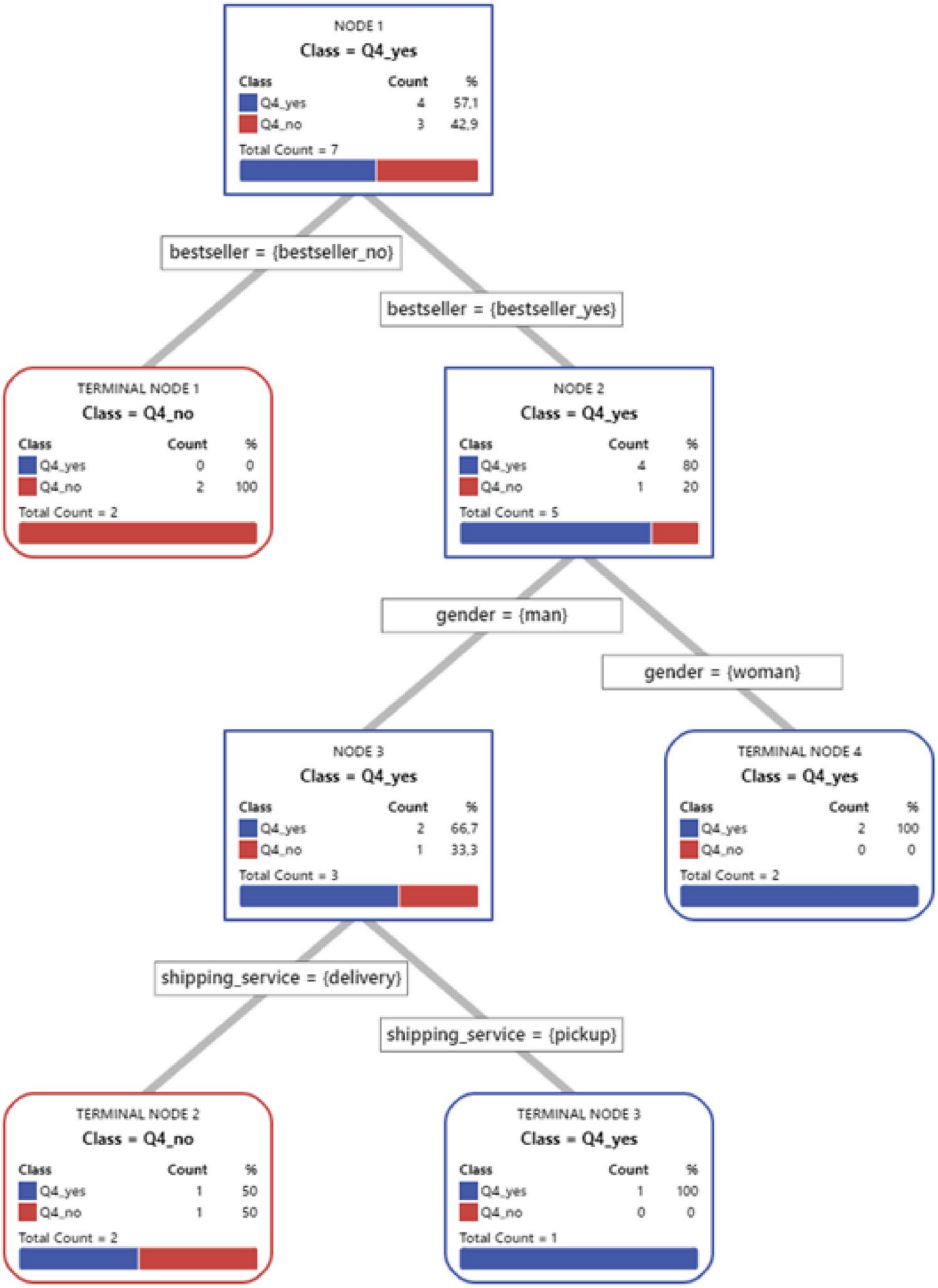

Fig. 1 Decision tree in our running example [Node 1 represents the root of the decision tree with all orders (transactions) from the training set *D*. Node 2 and Node 3 represent the internal nodes of the decision tree after the first and the second recursive call of Algorithm 1. Terminal nodes 1–4 represent the leaves of the decision tree after satisfying the stopping criteria. Each node highlights the mode of the target attribute and the absolute (or relative) frequencies of the categories (classes). The edges include the values of the splitting attribute]

bestseller attribute. We recursively call the method for decision tree construction on these two subsets of orders (i.e., bestsellers and not bestsellers.)

The decision tree in our running example (Fig. 1) contains six nodes, whereby the most important input attributes (in three iterations of the algorithm) are the bestseller, the gender, and the shipping service, respectively.

Decision trees may have issues with overfitting, bias toward attributes with many categories, or sensitivity to small changes in the data. We address these limitations and discuss potential solutions in the discussion and limitations part of our paper.

In our paper, we apply the construction of a decision tree (Algorithm 1) for the classification of 9123 orders having the set of input attributes $A = \{$gross_profit_category,gender,country, shipping_service, bestseller, long_term_unavailable$\}$, and the target attribute $b =$ Q4_period such that

$L_{\text{gross\_profit\_category}} = \{\text{A,B,C}\}$,

$L_{\text{gender}} = \{\text{woman,man}\}$,

$L_{\text{country}} = \{\text{BG,CZ, DE, HR, HU, RO, SK,UA}\}$,

$L_{\text{shipping\_service}} = \{\text{delivery,pickup}\}$,

$L_{\text{bestseller}} = \{\text{yes,no}\}$,

$L_{\text{long\_term\_unavailable}} = \{\text{yes,no}\}$,

$L_{\text{Q4\_period}} = \{\text{Q4\_yes,Q4\_no}\}$.

GUHA association rules

GUHA (acronym for General Unary Hypotheses Automaton) is a method for generating the association rules between attributes in the data. GUHA methods were successfully applied in the analysis of patient examination (e.g., for attributes of cholesterol, weight, height, triglycerides, and other), for analysis of forest damage (e.g., the effects of sulfur or chlorine on needle loss of spruce), prediction of highway travel time, mutagenes discovery, determination of suitable markers for the prediction of bleeding at patients with chronic lymphoblastic leukemia, or analysis of data about epileptic patients (Hájek et al. 2010).

In comparison with the Apriori algorithm (Agrawal et al. 1993), GUHA method provides a much more general set of association rules. In particular, one can generate association rules based on different types of associations between the attributes and different logical formulas. GUHA method is based on the so-called interest measures or quantifiers (for example p-implication) which we describe in the following.

Consider a set of attributes A and a value $0 < p \leq 1$. Then, a p-implication over A is an expression

$$Z \Rightarrow_p W,$$

whereby Z is an antecedent, W is a consequent, and $Z, W \subseteq A$. The p-implication is true if at least $100 \cdot p\%$ of objects satisfying Z fulfills W, as well.

	Gender	Shipping_service	Bestseller	Period
Order #36	Man	Pickup	No	Q4_no
Order #47	Woman	Delivery	No	Q4_no
Order #128	Man	Delivery	Yes	Q4_no
Order #151	Woman	Delivery	Yes	Q4_yes
Order #160	Man	Delivery	Yes	Q4_yes
Order #199	Woman	Pickup	Yes	Q4_yes

In our example, the 1-implication can be illustrated by the expression $\{\text{woman, bestseller\_yes}\} \Rightarrow_1 \{\text{Q4\_yes}\}$ which means that women with bestseller ordered their products in the fourth quarter of the year. Moreover, $p = 1$ implies it holds for at least 100% of orders in our data.

Now, we describe how to determine whether p-implication is true in the dataset. Let D be a training set. Thus, for two binary attributes $a_1, a_2 \in A$, we can construct the contingency table, whereby

- r is the number of objects of a training set D satisfying both a_1 and a_2,
- s is the number of objects of a training set D satisfying both a_1 and $\neg a_2$,
- t is the number of objects of a training set D satisfying both $\neg a_1$ and a_2,
- u is the number of objects of a training set D satisfying both $\neg a_1$ and $\neg a_2$.

D	a_2	$\neg a_2$
a_1	r	s
$\neg a_1$	t	u

In our running example, we can take a_1 = bestseller and a_2 = Q4_period. Thus, we obtain

	Q4_yes	Q4_no
bestseller_yes	3	1
bestseller_no	0	2

In general, p-implication

$$\{a_1\} \Rightarrow_p \{a_2\}$$

is true in D if

$$\frac{r}{r+s} \geq p.$$

In our running example, we have that 0.75-implication

$$\{\text{bestseller\_yes}\} \Rightarrow_{0.75} \{\text{Q4\_yes}\}$$

is true in D since

$$\frac{3}{3+1} \geq 0.75.$$

However, 1-implication

$$\{\text{bestseller\_yes}\} \Rightarrow_1 \{\text{Q4\_yes}\}$$

is not true in D since

$$\frac{3}{3+1} \ngeq 1.$$

Generally, p-implications can be seen as the association rules from Agrawal et al. (1993) with confidence p. To explain how GUHA association rules differ from other association rule mining techniques and to understand its applicability and potential advantages, we present the additional types of rules that can be generated by GUHA method. In particular, we can use GUHA to generate the so-called double p-implication over a set of attribute A. The double p-implication over A is a formula

$$Z \Leftrightarrow_p W,$$

that is true if at least $100 \cdot p\%$ of objects (e.g., customers) satisfying W or Z fulfill W and Z simultaneously. Double p-implication

$$\{a_1\} \Leftrightarrow_p \{a_2\}$$

is true in D if

$$\frac{r}{r+s+t} \geq p.$$

In our running example, we have that double 0.75-implication

$$\{\text{bestseller\_yes}\} \Leftrightarrow_{0.75} \{\text{Q4\_yes}\}$$

is true in D since

$$\frac{3}{3+1+0} \geq 0.75.$$

However, 1-implication

$$\{\text{bestseller\_yes}\} \Leftrightarrow_1 \{\text{Q4\_yes}\}$$

is not true in D since

$$\frac{3}{3+1+0} \ngeq 1.$$

The third possibility is constructing p-equivalence over a set of attributes A. In particular, p-equivalence over A is a formula

$$Z \equiv_p W,$$

that is true if at least $100 \cdot p\%$ of all objects (e.g., customers) obtain the same truth degree for Z and W. Thus, p-equivalence

$$\{a_1\} \equiv_p \{a_2\}$$

is true in D if

$$\frac{r+u}{r+s+t+u} \geq p.$$

In our running example, we have that 0.75-equivalence

$$\{\text{bestseller\_yes}\} \equiv_{0.75} \{\text{Q4\_yes}\}$$

is true in D since

$$\frac{3+2}{3+1+0+2} \geq 0.75.$$

However, 1-equivalence

$$\{\text{bestseller\_yes}\} \equiv_1 \{\text{Q4\_yes}\}$$

is not true in D since

$$\frac{3+2}{3+1+0+2} \ngeq 1.$$

The other GUHA measures include, for example, average dependence, simple deviation, Fisher quantifier, or above-average dependence (Hájek et al. 2010).

We explained the process of evaluation of p-implications for two binary attributes. Nevertheless, it can be generalized for more than two attributes. In this case, we will consider that r is the number of objects of a training set D satisfying all attributes from an antecedent and all attributes from a consequent. The same holds for computing s, t, and u in case of more than two attributes. Moreover, we can generate the contingency tables for categorical attributes by applying their transformations to the binary attributes.

In our running example, we can take the bestseller, shipping service, and Q4 period attributes. Thus, we obtain

	Q4_yes	Q4_no
bestseller_yes ∧ delivery	2	1

In our running example, we have that 0.75-implication

$$\{\text{bestseller\_yes, delivery}\} \Rightarrow_{0.6} \{\text{Q4\_yes}\}$$

is true in D since

$$\frac{2}{2+1} \geq 0.6.$$

However, 0.75-implication

$$\{\text{bestseller\_yes}\} \Rightarrow_{0.75} \{\text{Q4\_yes}\}$$

is not true in D since

$$\frac{2}{2+1} \ngeq 0.75.$$

In addition, the ratio of the number of objects in D with attributes from antecedent and consequent to the number of all objects in D is called the support of the rule. In our example, the support of 0.75-implication

$$\{\text{bestseller\_yes, delivery}\} \Rightarrow_{0.6} \{\text{Q4\_yes}\}$$

is 0.33 since we have two orders satisfying this rule out of all six orders.

In this paper, we will generate GUHA association rules for four categorical attributes of shipping_service, gender, gross_profit_category, and category_ID at antecedent and three categorical attributes of day_part, quarter_of_year, and day_name at consequent.

Formal concept analysis

Regarding data analysis methods, FCA is a bi-clustering method based on clustering the rows and columns of a table matrix. It provides powerful computational and visualization techniques for solving the various issues of supervised and unsupervised learning tasks (Ganter and Wille 1999). The connections of FCA with machine learning methods and their potential applications in text mining, biology, medicine, or bioinformatics are reviewed by Carpineto and Romano (2004); Poelmans et al. (2013).

We will describe the application of FCA with an example given by a small subset of our data. As the **input** of this method, we consider six fitness store orders described by five binary attributes (woman, man, delivery, pickup, and bestseller). Each order (or its attribute) can be represented by one row (or column) of the object-attribute table.

	Woman	Man	Delivery	Pickup	Bestseller
Order #11		×	×		×
Order #23		×	×		
Order #39		×		×	
Order #55		×		×	
Order #135	×			×	
Order #156	×			×	

Note that this cross table is generally rectangular in which the rows represent objects and the columns correspond to the attributes.

We can generally describe the cross object-attribute table by defining a notion of the so-called **formal context**.

Definition 1 Consider the non-empty sets B, A, and a crisp binary relation $I \subseteq B \times A$. A triple $\langle B, A, I \rangle$ is called a formal context. The elements of the sets B, A are called objects and attributes, respectively. A relation I is called an incidence relation.

In our example, we have that

$$B = \{\mathrm{order\#11}, \mathrm{order\#23}, \mathrm{order\#39}, \mathrm{order\#55}, \mathrm{order\#135}, \mathrm{order\#156}\}$$

and

$$A = \{\mathrm{woman{,}man{,}delivery{,}pickup{,}bestseller}\}.$$

Moreover, it holds that

$$\langle \mathrm{order\#11}, \mathrm{woman} \rangle \notin I,$$
$$\langle \mathrm{order\#11}, \mathrm{pickup} \rangle \notin I,$$

but

$$\langle \mathrm{order\#11}, \mathrm{man} \rangle \in I,$$
$$\langle \mathrm{order\#11}, \mathrm{delivery} \rangle \in I,$$
$$\langle \mathrm{order\#11}, \mathrm{bestseller} \rangle \in I.$$

Regarding the formal context of fitness shop orders, we aim to discover and visualize the meaningful biclusters in our data. For example, we are interested in only the maximal subsets of orders with a delivery method and the bestsellers. For such reasoning, we will use two natural and symmetric mappings to assign the common attributes of a given subset of objects and vice versa.

Definition 2 Let $\langle B, A, I \rangle$ be a formal context. Consider the subsets $X \in \mathrm{P}(B)$, $Y \in \mathrm{P}(A)$ of power sets of objects and attributes, respectively. Two mappings $\nearrow : \mathrm{P}(B) \to \mathrm{P}(A)$ and $\swarrow : \mathrm{P}(A) \to \mathrm{P}(B)$ such that

$$\nearrow (X) = X^{\nearrow} = \{y \in A : (\forall x \in X) \langle x, y \rangle \in I\}$$
$$\swarrow (Y) = Y^{\swarrow} = \{x \in B : (\forall y \in Y) \langle x, y \rangle \in I\}$$

are called concept-forming operators of a formal context $\langle B, A, I \rangle$.

In our example, it holds

$$\nearrow (\{\mathrm{order\#23}\}) = \{\mathrm{man{,}delivery}\},$$

but

$$\swarrow (\{\mathrm{man{,}delivery}\}) = \{\mathrm{order\#11}, \mathrm{order\#23}\}.$$

The mapping $\nearrow$ assigns the common attributes of selected fitness shop orders (e.g., man and delivery method to the order #23). Symmetrically, the mapping $\swarrow$ assigns the subset of all fitness shop orders to the selected attributes (e.g., man and delivery method).

The previous definition of concept-forming operators provides a way to construct the so-called **formal concepts** (which can be seen as the maximal biclusters from the clustering point of view).

Definition 3 Consider a formal context $\langle B, A, I \rangle$ and the concept-forming operators $\nearrow$, $\swarrow$. Consider the subsets of objects and attributes $X \in \mathrm{P}(B)$, $Y \in \mathrm{P}(A)$. A pair $\langle X, Y \rangle$ such that it holds $\nearrow (X) = Y$, $\swarrow (Y) = X$ is called a formal concept of $\langle B, A, I \rangle$.

In our running example, we can take

$$X = \{\mathrm{order\#11}, \mathrm{order\#23}\}$$

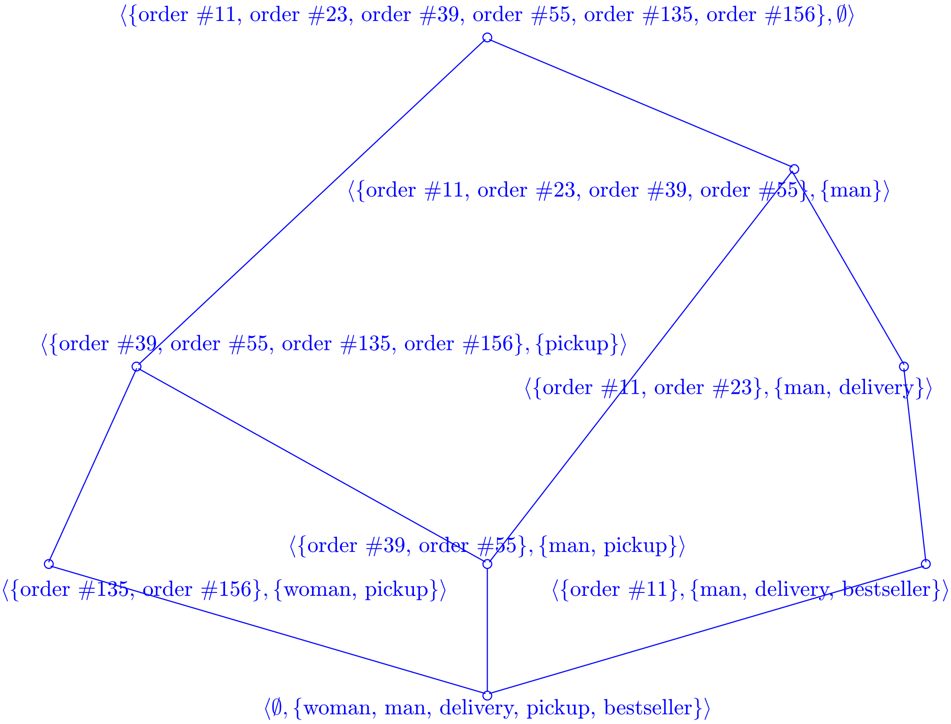
⟨{order #11, order #23, order #39, order #55, order #135, order #156}, ∅⟩
⟨{order #11, order #23, order #39, order #55}, {man}⟩
⟨{order #39, order #55, order #135, order #156}, {pickup}⟩
⟨{order #11, order #23}, {man, delivery}⟩
⟨{order #39, order #55}, {man, pickup}⟩
⟨{order #135, order #156}, {woman, pickup}⟩
⟨{order #11}, {man, delivery, bestseller}⟩
⟨∅, {woman, man, delivery, pickup, bestseller}⟩

Fig. 2 Concept lattice in our running example [Each node corresponds to the formal concept that includes the subset of objects (orders) and the subset of attributes. The concept lattice includes eight formal concepts, which are partially ordered. The edges represent the relationships between two formal concepts based on subsets of objects/attributes]

and

$$Y = \{\text{man,delivery}\}.$$

We have that

$$\nearrow (X) = \nearrow (\{\text{order\#11}, \text{order\#23}\}) = \{\text{man,delivery}\} = Y$$

and

$$\swarrow (Y) = \swarrow (\{\text{man,delivery}\}) = \{\text{order\#11}, \text{order\#23}\} = X.$$

Thus, a pair

$$\langle \{\text{order\#11}, \text{order\#23}\}, \{\text{man,delivery}\} \rangle$$

forms a formal concept.

We can imagine a formal concept as a bi-cluster that gathers a closed subset of objects and a closed subset of attributes. Another example of a formal concept can include the maximal subset of women who ordered with the pickup method.

In general, the set of all formal concepts (biclusters) can be expressed by a set

$$\{\langle X, Y \rangle \in \mathrm{P}(B) \times \mathrm{P}(A) : \nearrow (X) = Y, \swarrow (Y) = X\}.$$

In our running example, we obtain the following eight formal concepts:

$$\langle \emptyset, \{\text{woman,man,delivery,pickup,bestseller}\} \rangle,$$
$$\langle \{\text{order\#11}\}, \{\text{man,delivery,bestseller}\} \rangle,$$
$$\langle \{\text{order\#11}, \text{order\#23}\}, \{\text{man,delivery}\} \rangle,$$
$$\langle \{\text{order\#39}, \text{order\#55}\}, \{\text{man,pickup}\} \rangle,$$
$$\langle \{\text{order\#135}, \text{order\#156}\}, \{\text{woman,pickup}\} \rangle,$$
$$\langle \{\text{order\#39}, \text{order\#55}, \text{order\#135}, \text{order\#156}\}, \{\text{pickup}\} \rangle,$$
$$\langle \{\text{order\#11}, \text{order\#23}, \text{order\#39}, \text{order\#55}\}, \{\text{man}\} \rangle,$$
$$\langle \{\text{order\#11}, \text{order\#23}, \text{order\#39}, \text{order\#55}, \text{order\#135}, \text{order\#156}\}, \emptyset \rangle,$$

The last formal concept describes that all six orders have no common attribute. Symmetrically, there is no order with all attributes simultaneously (see first formal concept). However, we can see other meaningful closed groups of orders with common attributes.

Moreover, the set of all formal concepts of a formal context $\langle B, A, R \rangle$ can be ordered by a partial ordering relation. The partially ordered set of all formal concepts is called the concept lattice. This provides a way to construct the hierarchy of formal concepts and visualize the relationships between the formal concepts. The concept lattice in our running example is shown in Fig. 2.

However, we include the isomorphic visualization of a concept lattice with the so-called reduced labeling of eight formal concepts from our running example (Fig. 3). For better readability, only the first appearance of each object from the bottom (by white label) and the first appearance of each

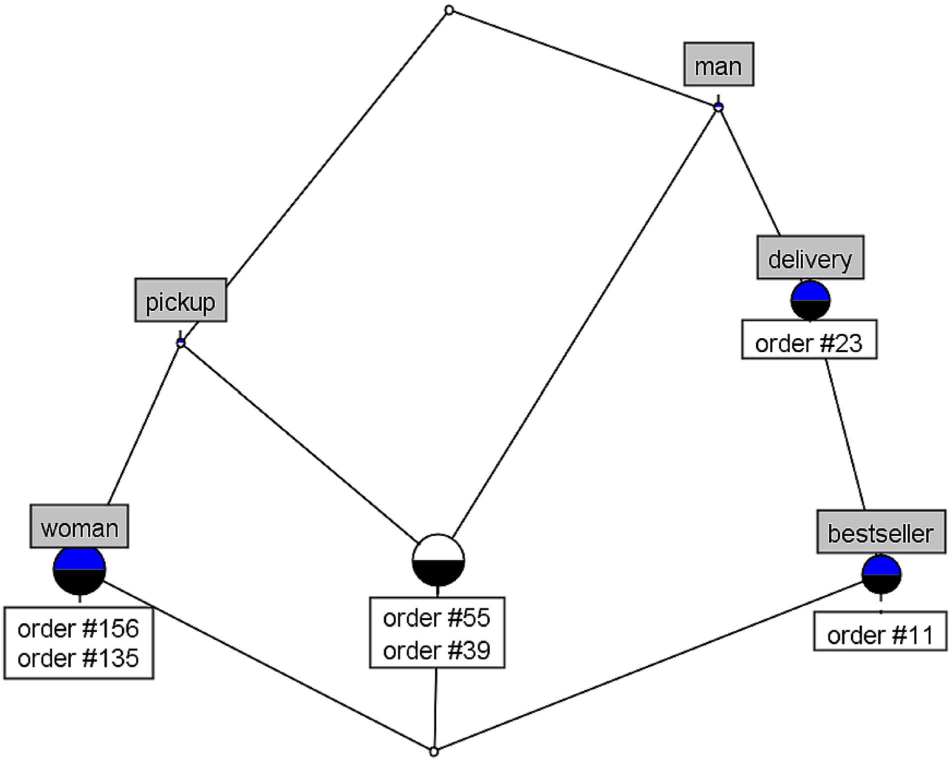

Fig. 3 Concept lattice with reduced labeling in our running example [The concept lattice includes eight formal concepts. Only the first appearance of each object from the bottom (by white label) and the first appearance of each attribute from the top (by shaded label) are highlighted. The attributes (objects) representing the selected formal concept can be obtained by collecting the shaded (white) labels leading to the top (to the bottom) of the concept lattice]

attribute from the top (by shaded label) are highlighted in the concept lattice with reduced labeling.

A concept lattice with reduced labeling illustrates the relationships between the subsets of orders regarding their common attributes. Consider white label with order #11 in the previous diagram. It represents the formal concept

$$\langle \{\text{order\#11}\}, \{\text{man,delivery,bestseller}\} \rangle.$$

If we select the white label with order #39, it represents the formal concept

$$\langle \{\text{order\#39, order\#55}\}, \{\text{man,pickup}\} \rangle.$$

Thus, the attributes representing the selected order can be obtained by collecting the shaded labels leading to the top of the concept lattice.

If we consider the shaded label with the attribute man, it represents the formal concept

$$\langle \{\text{order\#11, order\#23, order\#39, order\#55}\}, \{\text{man}\} \rangle.$$

Thus, if we select the shaded label of the attribute, the objects representing the selected attribute can be obtained by collecting the white labels leading to the bottom of the concept lattice.

Moreover, the relationships between subsets of orders can be linear or ordered by a partial ordering relation. For example, consider the order #11 with a delivery and the order number #39 with a pickup. They have a common attribute of man, but we cannot compare them linearly regarding the delivery method. Thus, they have a distance of one from the bottom element of the concept lattice with reduced labeling.

We will present the other concept lattices from our dataset in the following section devoted to our results. Note that our full dataset can be used as the input of this method since the gender, country, delivery method, and other attributes can be transformed into binary attributes. Moreover, FCA can provide another computational and visualization technique compared to decision trees and GUHA association rules.

In our paper, we applied FCA for two particular products of the fitness shop. The concept lattice of the first product is constructed concerning women's orders. The second concept lattice will correspond to the gender, dayparts, and delivery method.

Results

This research has focused on analyzing customer behavior related to the classification of orders by time aspects (decision trees), to the dependencies between attributes (GUHA method), and the biclusters of customers given by ID of product (FCA). We present the results of this research in this section.

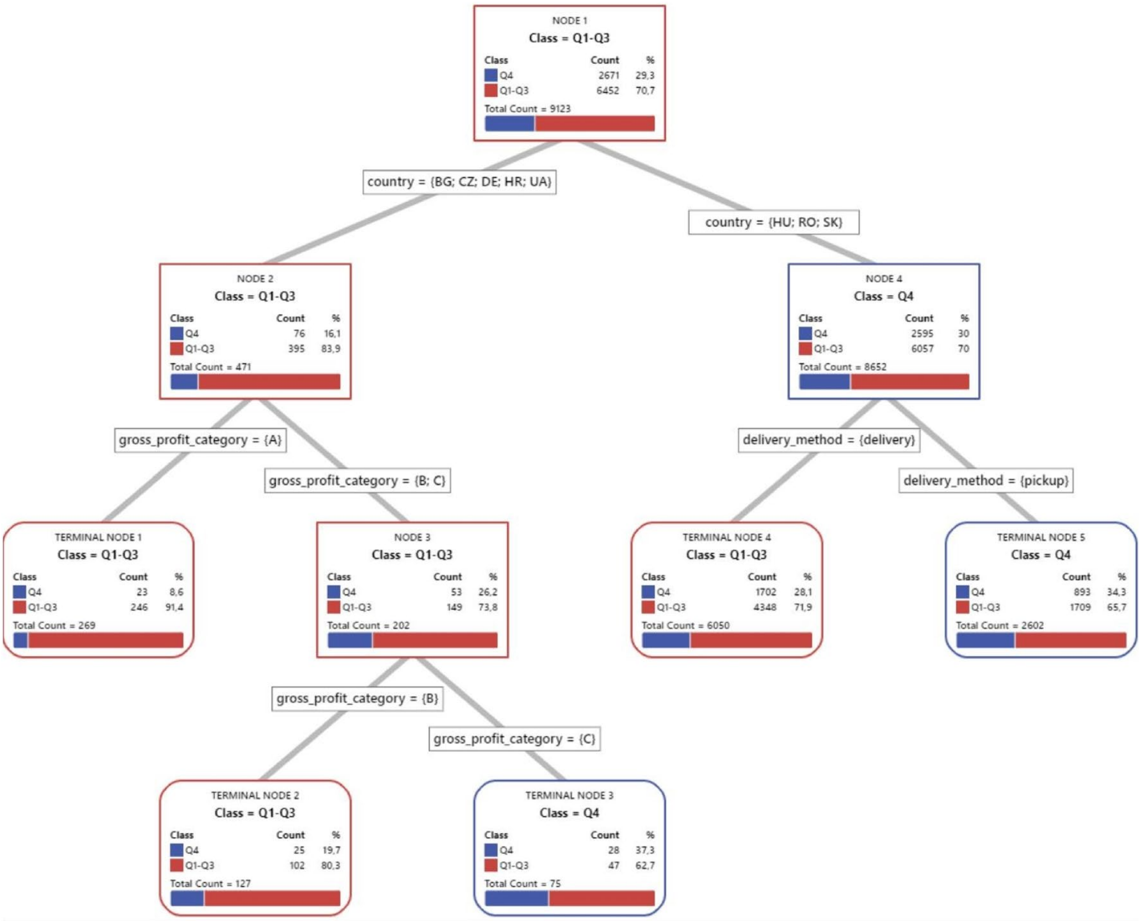

Fig. 4 Decision tree for classification of 9123 orders by Q4 period [Country is the input attribute with the best splitting quality (based on entropy). The decision tree terminates in five terminal nodes, which determine the paths from the root of the decision tree regarding at least 62% of orders in Q4 (or Q1–Q3) classes]

Results of decision trees

In the first part, we present the results obtained by a method of decision tree construction (Algorithm 1) implemented in Minitab tool. We used the entropy criterion for splitting or selecting attributes at each tree level. In particular, we obtained a decision tree illustrated by Fig. 4 for $A = \{\text{gross\_profit\_category,gender, country, delivery\_method,global\_bestseller,long\_term\_unavailable}\}$, and a binary target attribute $b = \text{Q4\_period}$.

The cross-validation with 5-fold was performed to assess the generalizability and reliability of the decision tree models. The accuracy of the decision tree was 77.6%. We found that 29.3% of customers ordered their products in the fourth quarter of the year (the root of the decision tree in Fig. 4). However, the proportion is only 16.1% if we select the customers from Bulgaria, Czech Republic, Germany, Croatia, and Ukraine, (left part of the second level of nodes of the decision tree in Fig. 4). Moreover, if the gross profit category of the product equals A, the proportion of orders in the fourth quarter is 8.6% (left part of the third level of nodes of the decision tree in Fig. 4). On the contrary, customers from Slovakia, Hungary, and Romania ordered 30% of their products in the fourth quarter of the year (right part of the second level of nodes of the decision tree in Fig. 4). If they selected the pickup method of delivery, their proportion is 34.3% (right part of the third level of nodes of the decision tree in Fig. 4).

We present four levels of nodes in our decision tree from Fig. 4 (including its root). However, additional levels were generated, as well. In the fifth level, we found that for customers from Bulgaria, Czech Republic, Germany, Croatia, and Ukraine, the gross profit category of B and long-term unavailable products, the orders were not booked in the fourth quarter (which is not shown in Fig. 4) due to lack of space and readability of decision tree in a graphical form.

The structure of the decision tree of the fourth quarter of the year implies that the most important attributes are country, gross profit category, and delivery method. Thus, the potential marketing strategies can include heterogeneous

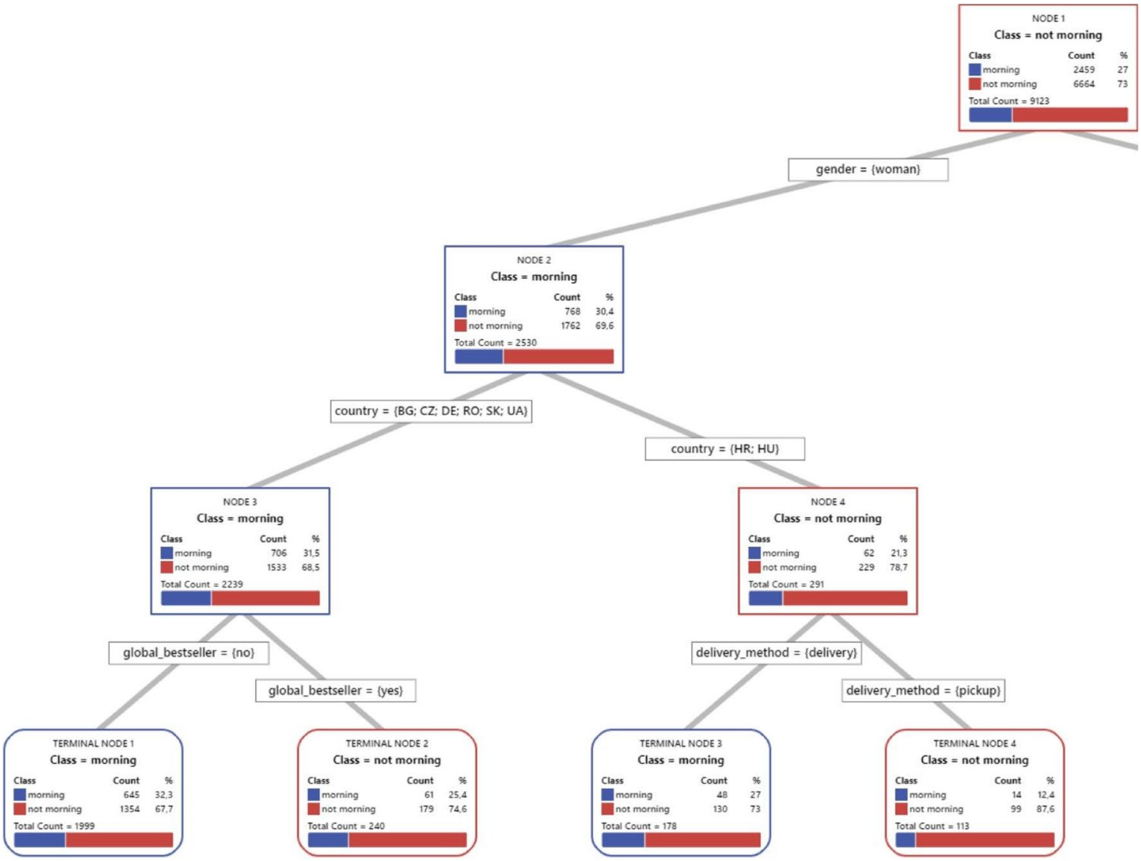

Fig. 5 Left part of decision tree for classification of 9123 orders by morning time [Gender is the input attribute with the best splitting quality (based on entropy). The left part of the decision tree determines the paths with women's orders regarding at least 67% of orders in the morning or not in the morning class]

conditions for each country based on its representatives. The observed patterns can be applied to a larger population or other regions to find additional hidden patterns in data.

We present the second decision tree of the same set of input attributes and a binary target attribute $b =$ morning in Figs. 5 and 6. The cross-validation with 5-fold was performed to assess the generalizability and reliability of the decision tree models. The accuracy of the decision tree was 71.2%. We found that 27% of customers ordered their products in the morning (first level of the decision tree in Figs. 5 and 6). The proportion of morning orders is 30.4% for women (second level of the decision tree in Figs. 5 and 6). However, if we consider women from Croatia and Hungary with the pickup method of delivery, we have 12.4% of orders in the morning.

The structure of the decision tree of the morning time implies that the most important attributes are gender and country (the first two levels of the decision tree). Thus, the potential marketing strategies can include heterogeneous conditions for men and women based on these findings. The observed patterns can be applied to a larger population or other regions to find additional hidden patterns in data.

Results of GUHA method

In Table 1, we present selected GUHA association rules for 4 categorical attributes of delivery_method, gender, gross_profit_category, and category_ID at antecedent and for 3 categorical attributes of day_part, quarter_of_year, and day_name at consequent. In particular, we obtained 250 association rules for 0.7-implications. Moreover, we selected the non-redundant GUHA association rules in Table 1. For example, the second GUHA association rule can be interpreted as follows. Our generated rules provide meaningful insights into customer preferences, patterns, or tendencies. For example, for woman with gross profit category A and product ID1923, the order was booked in the Saturday morning in the third quartal with 100% confidence.

Association rules typically include measures of confidence and support. Confidence indicates the conditional probability of the consequent given the antecedent, while support represents the frequency or prevalence of the rule in the dataset. The confidence and support of GUHA association rules are specified by the values of the last two columns in Table 1. Higher confidence values indicate stronger

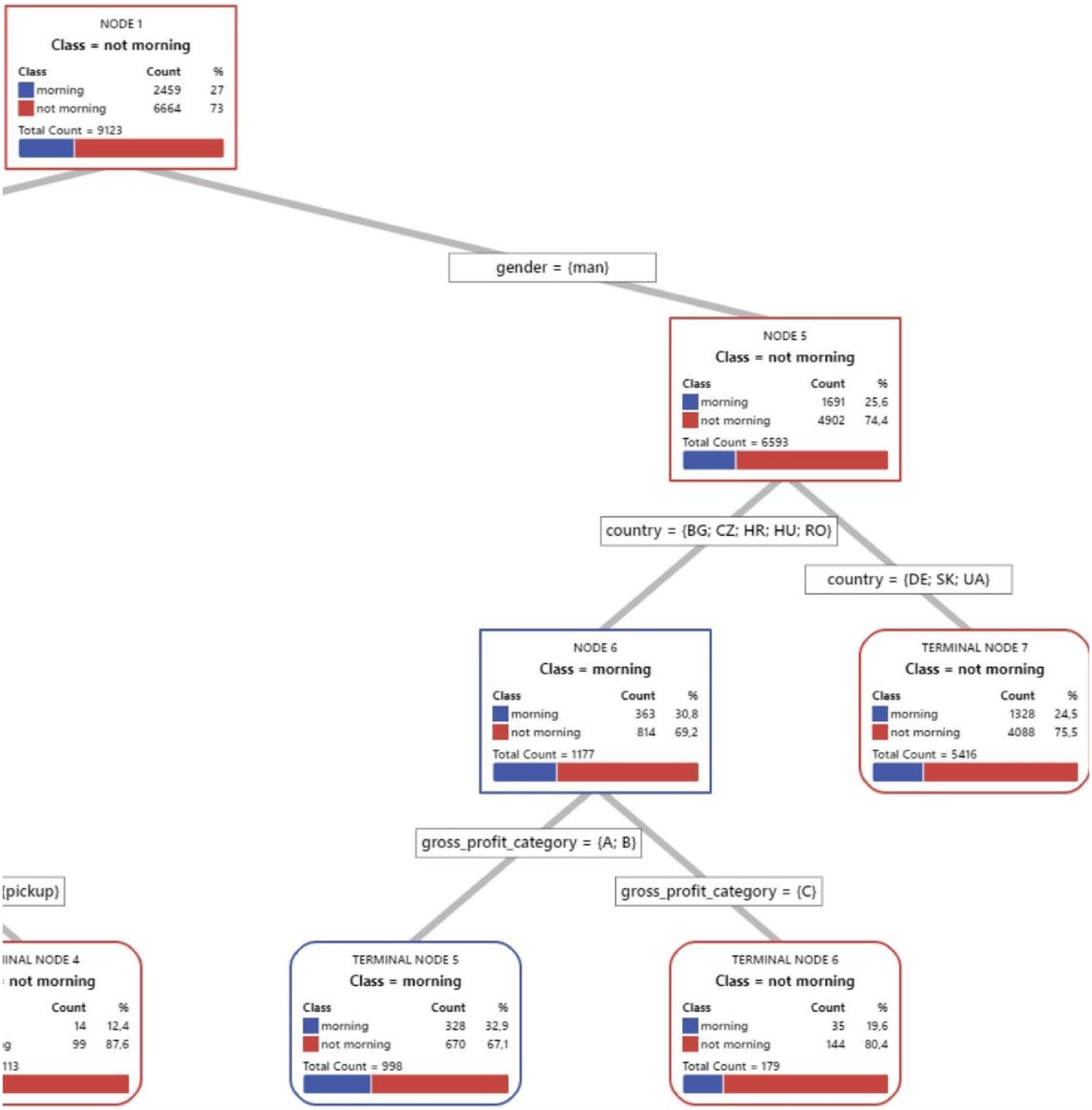

Fig. 6 Right part of decision tree for classification of 9123 orders by morning time [Gender is the input attribute with the best splitting quality (based on entropy). The right part of the decision tree determines the paths with men's orders regarding at least 67% of orders in the morning or not in the morning class]

associations, while higher support values suggest more frequent occurrences.

The rules in Table 1 were selected to avoid repetition and improve the association analysis's clarity and interpretability. The obtained association rules can be generalized to broader populations or contexts. In particular, the results show the typical period of day or week when men or women order particular products. The identified associations can be translated into actionable insights or strategies to improve marketing, sales, or customer service. The business can leverage the rules to enhance customer experiences, optimize product offerings, or tailor promotional campaigns based on the periods obtained by association rules.

Results of FCA

Based on FCA, we generated a formal context of 9123 objects (correspond to the orders) and 28 binary attributes. We transformed the original categorical attributes from our dataset into two binary attributes of gender, eight binary attributes of the country, two binary attributes of delivery method, three binary attributes of gross profit category, seven binary attributes of day name, and six binary attributes of day part.

In Fig. 7, we present the concept lattice with highlighted attributes for the product ID1923 ordered by women from Slovakia and Czech Republic with gross profit categories

Table 1 GUHA association rules for online customer behavior

No	GUHA association rule	*p*-imp	Supp
1	{delivery, man, B, product ID1791} $\Rightarrow_{0.7}$ {evening, Q1, Friday}	1	6
2	{woman, C, product ID2431} $\Rightarrow_{0.7}$ {evening, Q3, Sunday}	1	6
3	{delivery, woman, A, product ID1923} $\Rightarrow_{0.7}$ {morning, Q3, Saturday}	1	5
4	{delivery, man, A, product ID1912} $\Rightarrow_{0.7}$ {morning}	0.875	7
5	{delivery, woman, C, product ID1916} $\Rightarrow_{0.7}$ {early night, Q1, Wednesday}	0.867	13
6	{delivery, man, C, product ID1823} $\Rightarrow_{0.7}$ {lunch, Q2, Tuesday}	0.857	6
7	{pickup, woman, B, product ID1806} $\Rightarrow_{0.7}$ {lunch, Q4, Friday}	0.857	6
8	{woman, A, product ID2410} $\Rightarrow_{0.7}$ {Q4}	0.769	10
9	{pickup, woman, product ID1853} $\Rightarrow_{0.7}$ {Tuesday}	0.750	6
⋮	⋮	⋮	⋮

250 association rules were generated for 0.7-implications. The column with *p*-imp values expresses the conditional probability of the consequent given the antecedent. [Support represents the absolute frequency of the rule in the dataset

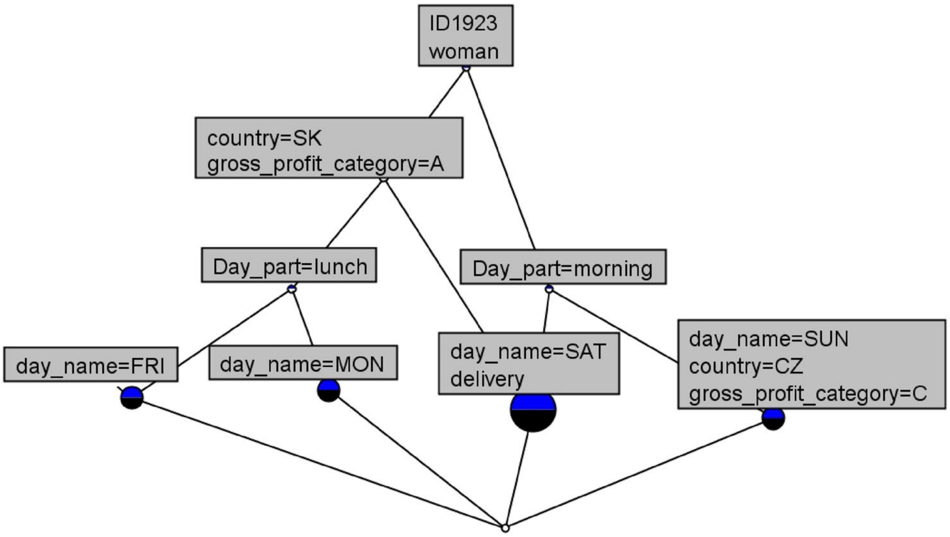

Fig. 7 Concept lattice with reduced labeling of product with ID1923 ordered by women [The concept lattice includes eight formal concepts. The shaded label highlights the first appearance of each attribute from the top of the concept lattice. The attributes representing the selected formal concept can be obtained by collecting the shaded labels leading to the top of the concept lattice. The size of black and blue nodes with one distance from the bottom of the concept lattice expresses the proportion of the orders in the particular node.]. (Color figure online)

A, C. The concept lattice is illustrated by ConExp[1] tool. This concept lattice includes nine nodes which represent nine formal concepts. The shaded labels of attributes ID1923, women linked to a node at the top of the concept lattice represent the objects we selected for our analysis. The shaded labels of attributes closer to the top element in the concept lattice are related to the attributes more often present in the objects. We can read the dependencies between attributes by the paths in the concept lattice. For example, the Friday and Monday lunch orders are from Slovakia and category A (left part of concept lattice in Fig. 7). The Sunday orders from Czech and category C were booked in the morning (right part of concept lattice in Fig. 7). Paths connecting attributes and objects indicate the presence of specific combinations or patterns. For example, in Fig. 7, the path from women to Friday lunch orders indicates that women are more likely to place orders on Fridays during lunchtime.

We present the concept lattice with highlighted objects and attributes of the product ID1912 in Fig. 8. Each object (i.e., white label in the concept lattice) is represented by the attributes (the shaded labels in the concept lattice) collected on all paths leading to the top of the concept lattice from the selected order node. The white labels (i.e., consumer orders) correspond to the groups of similar orders regarding the selected attributes. Each order and each attribute is shown in the concept lattice only once (so-called reduced labeling) for better readability. Three sets of orders which have distance one from the bottom of concept lattice (i.e., $\{9078, 8932, 8931, 8897, 8896, 8895, 8217, 8063, 7896\}$, $\{8237\}$, and $\{8804, 7881, 7880\}$) and represent three main groups of order representatives such that there is no order

[1] Concept Explorer, http://conexp.sourceforge.net.

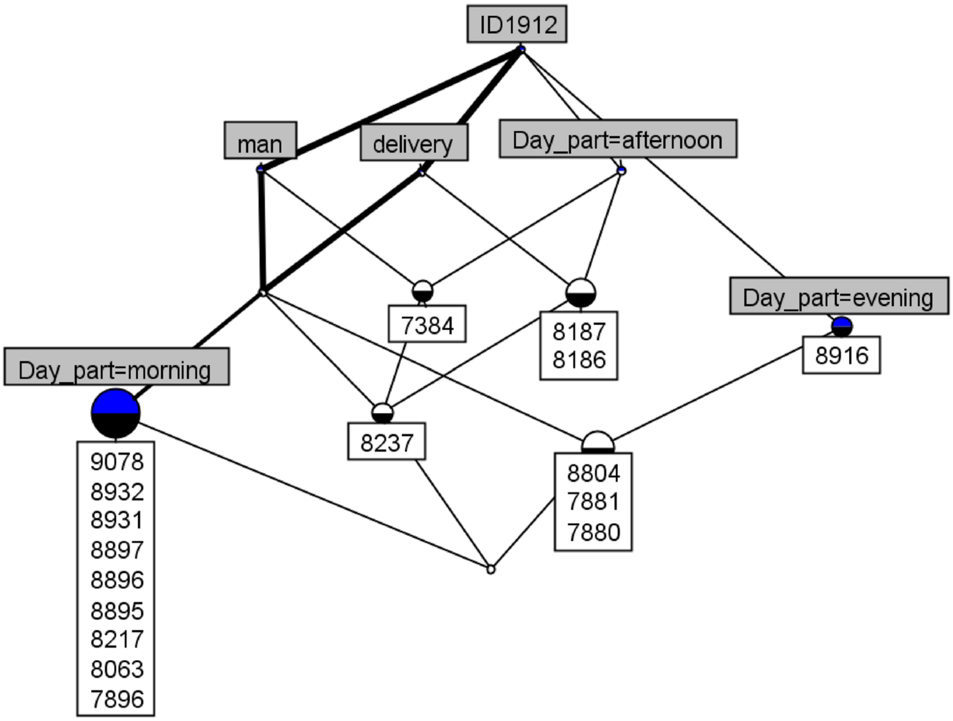

Fig. 8 Concept lattice with reduced labeling of product ID1912 [The concept lattice includes 12 formal concepts. The shaded label highlights the first appearance of each attribute from the top of the concept lattice. The attributes representing the selected formal concept can be obtained by collecting the shaded labels leading to the top of the concept lattice. The size of black nodes expresses the proportion of the orders in the particular node. Black and white nodes indicate no shaded attribute labels at these nodes]

with the superset of their attributes. For example, the group of orders {8804, 7881, 7880} is from men, classical delivery method, and evening since they are the attributes which are collected on all paths leading to the top of the concept lattice from the selected group. The path from the man to the morning indicates that men will most likely order this product in the morning. Analyzing these identified groups can be valuable for targeted marketing, personalized recommendations, or understanding customer preferences.

In addition, the orders covered by the particular attribute can be obtained by collecting the white labels leading to the bottom of the concept lattice from the selected attribute node. For example, we can see that afternoon is fulfilled by orders leading down from this node, i.e., {7384, 8187, 8186, 8237}. The thickness of edges represents the proportion of objects in the specific part of a concept lattice. The use of reduced labeling, highlighting, and shading can enhance the readability of the concept lattice and make it easier to interpret.

Discussion and limitations

Our current research has been accompanied by a few limitations that need to be taken into account in the application of the results.

Decision trees

Decision trees are recognized as a prominent technique in machine learning, primarily attributed to their inherent simplicity, transparency, and immediate efficacy. However, certain inherent constraints are their widespread adoption:

The propensity for overfitting

Particularly profound in deeper trees, there is a conspicuous tendency for overfitting, a scenario where the model inadvertently assimilates the noise from training data, undermining its potential to generalize effectively to unobserved data. Such concerns have been corroborated in numerous scholarly investigations. For example, an exploration by Garcia Leiva et al. (2019) underlined this persistent challenge, proffering an innovative method as a potential countermeasure (Garcia Leiva et al. 2019).

Extraction of attribute importance

Directly asking about the importance of attributes in decision trees can involve several drawbacks, especially in fields like customer satisfaction surveys (de Oña et al. 2016).

Predefined structure

Decision trees may struggle with learning the relations between labels without a predefined structure, which can lead to challenges in multi-label classification (Lotf and Rastegari 2020).

Evolutionary design

The evolutionary design of decision trees requires specialized approaches and recognized specializations, which can complicate the design process (Podgorelec et al. 2013).

GUHA method

GUHA method can be appropriate for finding dependencies between attributes, whereby the attributes are divided into two groups (antecedent and consequent). We can numerically evaluate the power of each rule based on several interest measures.

Scalability

An important issue of data analysis with GUHA method can be the high number of generated *p*-implications. Most of the generated implications are only consequences of already found patterns of domain knowledge. The simplification of rules or reduction of redundant association rules can be explored in this area. The selection of correct methods and parameters for dimensionality reduction can be challenging.

Interpretability

Interpretability can be seen as the degree to which people can understand the inducement of a decision. The high interpretability of GUHA methods can be demonstrated in various applications from a healthcare domain, medical data analysis, environmental issues, or travel time predictions.

Potential challenges

Real-world data often contain outliers or noise. Thus, the GUHA association rules can discover irrelevant implications. GUHA methods should be robust to noisy data. Moreover, missing data can cause the extraction of incomplete implications and results with bias.

FCA method

FCA is based on the theory of lattices, and it provides the construction of concept lattices with hierarchical dependencies between objects and attributes. It can be used for exploring specific products regarding the characteristics of customers or time aspects of orders. The advantage is the possibility to graphically illustrate a concept lattice.

Scalability

FCA is an appropriate method for datasets with one big dimension (objects or attributes) and a small or medium second dimension. However, several fruitful methods and approximations (e.g., the process of attribute exploration or pruning algorithms) can be used to reduce the computational complexity in the case of big datasets.

Interpretability

FCA provides a great visual tool for data interpretation. The precise formal definitions of input, operators, and output and the hierarchical structure of concept lattice contribute to its interpretability, making it a powerful technique for understanding heterogeneous data relationships and patterns in an intuitive way.

Potential challenges

One of the potential challenges in FCA is the construction of a formal context by the selection of the proper attributes and objects. If data are extracted from the database and the attributes are not systematically transformed, it can impact the results. However, several novel approaches and methods are proposed to reduce this potential challenge. Moreover, the area of fuzzy approaches in FCA has been thoroughly explored by several research groups in the world. Each of these approaches has advantages, allows to work with different mathematical structures, and can be applied for different tasks and applications.

Association rules

In academic discourse on data mining, association rules emerge as a widely recognized approach for discerning interrelations among database variables. Nevertheless, every methodology is tethered to its inherent drawbacks. The following encapsulates the prevalent limitations tied to association rules:

Volume of rules

Association rule algorithms can generate many rules, making it challenging to identify which ones are significant or beneficial (Agrawal et al. 1993).

Lack of causality

In academic discourse, it is imperative to clarify that association rules merely elucidate correlations and do not infer causation. The co-occurrence of two items with regularity does not necessarily imply a causal relationship between them (Tan et al. 2006).

Threshold dependency

The results heavily depend on the chosen threshold values for support and confidence. Selecting these parameters involves some subjectivity and can significantly influence the identified rules (Ghafari and Tjortjis 2019).

Static nature

Association rules are predominantly derived from static datasets. Consequently, their capacity to adapt effectively to evolving data or shifts in trends over temporal sequences may need to be improved (Wang et al. 1997).

Lack of context

These rules sometimes overlook the context of the transactions, which can lead to misleading associations that do not align with the true meaning of the data (Meruva and Bondu 2021).

Scalability issues

As the size of the dataset expands, the computational difficulty in generating association rules can become overwhelming. This issue is particularly acute without optimized algorithms or sufficient computational resources.

Limitation of research

Our study faces a challenge in accurately identifying the gender of each customer. The company employed two methods: a strict definition of gender for specific data and a predictive model for ambiguous cases. Although the model is generally reliable, with an 85–95% accuracy in 99% of instances, we cannot fully confirm its precision. However, we have accepted the model's predictions as conclusive for our research. Another issue is the inconsistency in name formats across purchases, as our company operates in various European markets, not all of which use the euro. For example, Hungary, the Czech Republic, and Romania use different currencies, necessitating the conversion of their orders to euros using the current exchange rate. This process, while necessary, introduces minor inaccuracies due to exchange rate fluctuations over time, particularly noticeable from 2014 to 2019. To mitigate this, the company should adopt a real-time conversion system that adjusts non-euro transactions to their euro equivalent at the prevailing rate, which would more effectively correct this limitation.

Recommendations

Analyzing consumers' online purchasing habits can provide many advantages for commercial entities, marketing professionals, and, indeed, the consumers themselves. The following elucidates the prospective benefits derived from such comprehension. As an example of potential implications for practice, we will talk about a fictional mid-sized e-commerce company wants to increase its market share and customer loyalty. The company has a diverse product range and a broad customer base but faces stiff competition from larger online retailers.

Personalized shopping experiences

A comprehensive analysis of a consumer's digital purchasing patterns enables enterprises to tailor shopping experiences to individual clients, suggesting products that align with their preferences.

Example of application

Objective To implement personalized marketing strategies that enhance customer engagement, increase conversion rates, and boost sales.

Data collection Collect data on customer demographics, past purchase history, browsing patterns, and engagement with previous marketing campaigns.

Data analysis Use the mentioned machine learning algorithms to segment customers into distinct groups based on their behavior and preferences.

Implementation Develop personalized email marketing campaigns that recommend products based on past purchases and browsing behavior.

Measurement Track key performance indicators (KPIs) such as open rates, click-through rates (CTRs), conversion rates, and average order value.

Improved marketing efficiency

Studying consumer behavior offers crucial insights to marketers, guiding them in allocating resources wisely, targeting specific customer groups accurately, and focusing on promoting particular products.

Example of application

Objective Optimize and streamline budget spending for marketing campaigns and improve performance metrics.

Data collection Based on the time stamp of the order, create new time variables using BI tools to help specify the time aspect of customer behavior more accurately.

Data analysis Using a combination of decision trees, GUHA association rules and especially FCA analysis, create a time profile of individual products within customer segments.

Implementation The company can enhance its marketing strategies by targeting ads more effectively at potential customers. The company can schedule ads strategically by understanding which products specific customer groups are inclined to purchase at certain times through customer behavior analysis. This precise timing for ad displays can significantly improve the cost-efficiency of their marketing spend.

Measurement To measure the cost-efficiency of a campaign, the company should consider several key metrics that can help you evaluate its performance: Cost Per Click (CPC), CTR, Conversion Rate or Return on Ad Spend (ROAS).

Predictive analytics is an area of sophisticated analytics that leverages past data, statistical algorithms, and machine learning methods to forecast future events. It aims not only to understand past events but also to deliver a well-informed estimate of future occurrences.

Example of application

Objective To utilize predictive analytics to forecast demand, optimize inventory levels, and personalize the shopping experience.
Data collection

- gather historical sales data, including seasonal variations and sales trends,
- track customer purchase patterns, including frequency, quantity, and types of products purchased,
- monitor external factors such as local events, holidays, and weather patterns that may influence shopping behavior

Data analysis

- use time-series analysis to understand sales trends and predict future demand for different product categories,
- apply machine learning models to identify patterns in customer purchase behavior and predict future buying trends,

*Implementation*Create personalized shopping experiences by predicting what products customers likely need based on their purchase history and predictive models.
Measurement

- monitor KPIs such as stock turnover rates, sell-through rates, markdown percentages, and customer satisfaction scores,
- continuously refine predictive models with new data and feedback to improve accuracy,
- use A/B testing to compare outcomes with and without predictive analytics to measure the impact.

Dynamic pricing strategies

Dynamic pricing strategies are a strategy where businesses set flexible prices for products or services based on current market demands. (Neubert, 2022) It is widely used in various industries, from hospitality to retail, and is particularly prevalent in online marketplaces.

Example of application

Objective Implement a real-time dynamic pricing strategy that adapts to market demand, competitor prices, and consumer buying trends.
Data collection Implement user behavior analytics tools to track how customers interact.
Data analysis

- gather historical sales data to understand how price changes affect demand,
- monitor competitor pricing for similar products in real time,
- track inventory levels to understand supply constraints,
- analyze customer data for price sensitivity.

Implementation Use predictive analytics to forecast demand for different products at various price points.
Measurement

- monitor sales volume, profit margins, and price competitiveness KPIs;
- continuously refine the pricing algorithm based on sales performance and market conditions.

Conclusion

In this paper, we presented the analysis of real e-shop consumer behavior data in eastern Europe, which contains data from 9123 orders of sports nutrition, healthy foods, fitness clothing, and accessories from 2014 to 2021. For analysis, we applied the methods of decision trees, GUHA association rules, and Formal concept analysis. We found that each of these methods provides various possibilities to have a greater look at consumer behavior of online e-shops.

In particular,

- *Decision trees* can help to classify the online customer orders with selected target attribute in data, e.g., time aspect or selected categories of customers.
- *GUHA method* can be appropriate for finding dependences between attributes of online customer behaviors, whereby the attributes are divided into two groups (antecedent and consequent).
- *FCA method* can help to explore and visualize the specific properties of products regarding the characteristics of customers or time aspects of orders.

Analyzing consumers' online purchasing habits can provide many advantages for commercial entities, marketing professionals, and consumers. Regarding the potential implications for practice, we discussed the examples of applications of personalized shopping experiences, improved marketing efficiency, predictive analytics, or dynamic pricing strategies.

In our future work, we aim to explore other machine learning methods of supervised learning (classification and regression) and unsupervised learning (clustering, attribute implications, or association rules) to analyze online consumer behavior.

Acknowledgements This article was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic under contract VEGA 1/0645/22 entitled by Proposal of Novel Methods in the Field of Formal Concept Analysis and their Application. The Slovak Research and Development Agency supported this work under the Contract No. APVV-20-0069, by the Scientific Grant Agency (VEGA) under Grant No. VEGA 2/0097/20, and by the Operational Programme Integrated Infrastructure (OPII) for the project 313011BWH2: "InoCHF" Research and development in the field of innovative technologies in the management of patients with "CHF," co-financed by the European Regional Development Fund.

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak Republic.

Data availability Data available on request from the authors.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Agrawal, R., T. Imieliński, and A. Swami. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 207–216. https://doi.org/10.1145/170036.170072.

Bellini, P., L.A.I. Palesi, P. Nesi, and G. Pantaleo. 2023. Multi clustering recommendation system for fashion retail. *Multimedia Tools and Applications* 82 (7): 9989–10016. https://doi.org/10.1007/s11042-021-11837-5.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24 (2): 123–140.

Carpineto, C., and G. Romano. 2004. *Concept data analysis. Theory and applications*. Chichester: Wiley.

Charandabi, S., and O. Ghanadiof. 2022. Evaluation of online markets considering trust and resilience: A framework for predicting customer behavior in e-commerce. *Journal of Business and Management Studies* 4 (1): 23–33. https://doi.org/10.32996/jbms.2022.4.1.4.

Charbuty, B., and A. Abdulazeez. 2021. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20 - 28. https://doi.org/10.38094/jastt20165.

de Oña, J., R. de Oña, and C. Garrido. 2016. Extraction of attribute importance from satisfaction surveys with data mining techniques: A comparison between neural networks and decision trees. *Transportation Letters: The International Journal of Transportation Research* 9 (1): 39–48. https://doi.org/10.1080/19427867.2015.1136917.

Fan, B., J. Yao, D. Lei, and R. Tong. 2022. Representation, mining and analysis of unsafe behaviour based on pan-scene data. *Journal of Thermal Analysis and Calorimetry* 148: 5071–5087 (2023). https://doi.org/10.1007/s10973-022-11655-3

Ganar, C., and P. Hosein. 2022. Customer segmentation for improving marketing campaigns in the banking industry. In 2022 5th Asia conference on machine learning and computing (ACMLC), 48–52. https://doi.org/10.1109/ACMLC58173.2022.00017.

Ganter, B., and R. Wille. 1999. *Formal concept analysis: Mathematical foundations*. Berlin: Springer.

Garcia Leiva, R., A. Fernandez Anta, V. Mancuso, and P. Casari. 2019. A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *IEEE Access* 7: 99978–99987. https://doi.org/10.48550/arXiv.1906.01246.

Ghafari, S.M., and C. Tjortjis. 2019. A survey on association rules mining using heuristics. *WIREs Data Mining and Knowledge Discovery* 9: e1307. https://doi.org/10.1002/widm.1307.

Hájek, P., M. Holeňa, and J. Rauch. 2010. The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences* 76 (1): 34–48. https://doi.org/10.1016/j.jcss.2009.05.004.

Husein, A.M., D. Setiawan, A.R.K. Sumangunsong, A. Simatupang, and S.A. Yasmin. 2022. Combination grouping techniques and association rules for marketing analysis based customer segmentation. *Sinkron Jurnal Dan Penelitian Teknik Informatika* 7 (3): 1998–2007. https://doi.org/10.33395/sinkron.v7i3.11571.

Kwon, S.E., Y.T. Kim, H. Suh, and H. Lee. 2023. Identifying the mobile application repertoire based on weighted formal concept analysis. *Expert Systems with Applications* 173: 114678. https://doi.org/10.1016/j.eswa.2021.114678.

Lotf, A., and R. Rastegari. 2020. Multi-label classification: A novel approach using decision trees for learning label-relations and preventing cyclical dependencies: Relations Recognition and Removing Cycles (3RC). In SITA'20: Proceedings of the 13th international conference on intelligent systems: Theories and applications. https://doi.org/10.1145/3419604.3419763

Luo, R. 2023. Predicting and visualization analysis of customer churn in telecommunications leveraging decision tree model. *Journal of Communication and Computer* 17: 3938. https://doi.org/10.54254/2755-2721/17/20230938.

Meruva, S.R., and V. Bondu. 2021. Review of association mining methods for the extraction of rules based on the frequency and utility factors. *International Journal of Information Technology Project Management (IJITPM)* 12 (4): 1–10. https://doi.org/10.4018/IJITPM.2021100101.

Plotnikova, V., M. Dumas, and F. Milani. 2021. Adapting the CRISP-DM data mining process: A case study in the financial services domain. In *Research challenges in information science*, vol. 415, ed. S. Cherfi, A. Perini, and S. Nurcan, 55–71. Cham: Springer. https://doi.org/10.1007/978-3-030-75018-3_4.

Podgorelec, V., M. Šprogar, and S. Pohorec. 2013. Evolutionary design of decision trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3 (2): 237–254. https://doi.org/10.1002/widm.1079.

Poelmans, J., D. I. Ignatov, S.O. Kuznetsov, and G. Dedene. 2013. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Application* 40 (16): 6538–6560. https://doi.org/10.1016/j.eswa.2013.05.009.

Quinlan, J.R. 1992. *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann. https://doi.org/10.1007/bf00993309.

Roscoe, S., M. Khatri, A. Voshall, S. Batra, S. Kaur, and J. Deogun. 2022. Formal concept analysis applications in bioinformatics. *ACM Computing Surveys*. https://doi.org/10.1145/3554728.

Roy, A. 2016. Relationship between consumers' purchase volume and purchase behaviour: A study on grocery buying in Kolkata. *Pacific Business Review International* 1 (4): 106–113. http://www.pbr.co.in/2016/2016_month/September/13.pdf.

Rungruang, Ch., P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub. 2024. RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications* 237 (Part B): 121449. https://doi.org/10.1016/j.eswa.2023.121449.

Saltz, J.S., and I. Krasteva. 2022. Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science* 8: e862. https://doi.org/10.7717/peerj-cs.862.

Samarth, V. 2023. *Understanding the Decision Tree: A guide to making better business decisions*. Emeritus. Accessed 31 Oct 2023. https://emeritus.org/in/learn/data-science-decision-tree/.

Schröer, C., F. Kruse, and J.M. Gómez. 2021. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science* 181: 526–534. https://doi.org/10.1016/j.procs.2021.01.199.

Stuti, S., K. Gupta, N. Srivastava, and A. Verma. 2022. A novel approach of product recommendation using utility-based association rules. *International Journal of Information Retrieval Research (IJIRR)* 12 (1): 1–19. https://doi.org/10.4018/IJIRR.289574.

Tan, P.N., M. Steinbach, and V. Kumar. 2006. *Introduction to data mining*. Indianapolis: Pearson Addison Wesley.

Tariq, A., J. Yan, A. Gagnon, M.R. Khan, and F. Mumtaz. 2022. Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geo-spatial Information Science*. https://doi.org/10.1080/10095020.2022.2100287.

Tundo, T., and F. Mahardika. 2023. Fuzzy inference system Tsukamoto-Decision Tree C 4.5 in predicting the amount of roof tile production in Kebumen. *JTAM (Jurnal Teori Dan Aplikasi Matematika)* 7 (2): 533. https://doi.org/10.31764/jtam.v7i2.13034.

Vaca, C., D. Riofrío, N. Pérez, and D. Benítez. 2020. Buy & sell trends analysis using decision trees. In *2020 IEEE Colombian conference on artificial intelligence (ColCACI)*, 1–6. IEEE. https://doi.org/10.1109/ColCACI50549.2020.9247907.

Wang, J., F. Shahzad, I. Khan, and A.W. Siyal. 2022. Decision making with an alternative mindset in an online shopping environment: Identifying user intentions toward Facebook-Commerce. *Frontiers in Psychology* 13: 1–17. https://doi.org/10.3389/fpsyg.2022.848931.

Wang, K., J. Yang, and R.R. Muntz. 1997. STING: A statistical information grid approach to spatial data mining. *VLDB* 7: 186–195.

Wen, Z. 2023. Feature analysis and model comparison of logistic regression and decision tree for customer churn prediction. *Journal of Communication and Computer* 20: 1073. https://doi.org/10.54254/2755-2721/20/20231073.

Xiao, B., and G. Piao. 2022. Analysis of influencing factors and enterprise strategy of online consumer behavior decision based on association rules and mobile computing. *Wireless Communications and Mobile Computing* 2022: Article ID 6849017. https://doi.org/10.1155/2022/6849017.

Yang, X., Z. Xu, and W. Zhou. 2019. Green credit product design based on fuzzy concept lattice. *IEEE Access* 7: 141574–141586. https://doi.org/10.1109/ACCESS.2019.2943486.

Yawata, K., Y. Osakabe, T. Okuyama, and A. Asahara. 2022. QUBO Decision Tree: Annealing machine extends decision tree splitting. In *2022 IEEE international conference on knowledge graph (ICKG)*, Orlando, FL, pp. 355–364. https://doi.org/10.1109/ICKG55886.2022.00052.

Zhou, L., L. Sun, Z. Li, W. Li, N. Cao, and R. Higgs. 2020. Study on a storage location strategy based on clustering and association algorithms. *Soft Computing*. 24 (3): 5499–5516. https://doi.org/10.1007/s00500-018-03702-9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Tomáš Pitka is an internal doctoral student at the Department of Applied Mathematics and Economic Informatics, EkF TU in Košice. He teaches subjects Informatics I, Introduction to Datamining. His areas of research are digital marketing, data analysis, and data mining using the help of innovative Business Intelligence tools in order to make processes and decision-making in companies more efficient.

Jozef Bucko is an Associate Professor at the Department of Applied Mathematics and Business Informatics, Faculty of Economics, Technical university of Košice, Slovakia. He defended his dissertation in the area of Graph Theory and habilitated in the area of Computer Science. His research focuses on digital systems security, information management, data processing, and search for behavioral patterns by applying machine learning methods and Business Intelligence tools in extensive digital data analysis.

Stanislav Krajči is a Full Professor at the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia. His research interests lie in the area of (fuzzy) logic, natural language processing, and Formal Concept Analysis. Recently, he contributed to the book Social Networks: A Framework of Computational Intelligence, and he is a reviewer of journals such as Fuzzy Sets and Systems, and Information Sciences.

Ondrej Krídlo is an Associate Professor at the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia. He received his PhD degree in Computer. He received the best paper award at the 10th International Conference on "Concept Lattices and their Applications" in La Rochelle, France (2013). He was a chair of the Organization Committee of the 11th International Conference on "Concept Lattices and their Applications" which was held in Ko?sice, Slovakia (2014).

Ján Guniš is an Associate Professor at the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia. He defended his dissertation in the area of the Theory of Teaching Informatics. Primarily, he is engaged in teaching programming, didactics of programming, diagnostics, and the development of computational thinking of pupils and students. He is a co-organizer of the Computer Science Teachers Club and the PALMA junior programming competition

Ľubomír Šnajder is an Associate Professor at the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia. He defended his dissertation in the area of the Theory of teaching mathematics and habilitated in the area of the Theory of teaching informatics. He specializes in diagnosing students' computer thinking and its development through programming, diagnosing students' programming misconceptions, and developing inquiry-based and project-based methodologies for teaching informatics. He is also co-organizer, author, and corrector of competition tasks of the PALMA junior programming competition in Python for secondary and primary school students.

Ľubomír Antoni is an Associate Professor at the Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovakia. He defended his dissertation in the area of Computer Science. He aims to propose selected machine learning methods and applications of these methods in solving the tasks of supervised and unsupervised learning. His scientific work is focused on methods of Formal Concept Analysis, fuzzy logic, and machine learning.

Peter Eliaš is a Scientific Researcher at the Mathematical Institute of Slovak Academy of Sciences in Košice, Slovakia. His research interests include general topology and real analysis. He is the co-investigator of projects entitled Probabilistic, Algebraic and Quantum Mechanical Methods of uncertainty determination or Algebraic and Topological Aspects of aggregation functions.