IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2021, accepted January 7, 2022, date of publication January 12, 2022, date of current version February 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3142537

A Novel Framework for Unification of Association Rule Mining, Online Analytical Processing and Statistical Reasoning

RAHUL SHARMA[1], (Graduate Student Member, IEEE), MINAKSHI KAUSHIK[1], SIJO ARAKKAL PEIOUS[1], ALEXANDRE BAZIN[2], SYED ATTIQUE SHAH[1], IZTOK FISTER, JR.[3], (Member, IEEE), SADOK BEN YAHIA[4], AND DIRK DRAHEIM[1]

[1]Information Systems Group, Tallinn University of Technology, 12616 Tallinn, Estonia
[2]Lorraine Research Laboratory in Computer Science and Its Applications (LORIA), CNRS, Inria, Université de Lorraine, 54000 Nancy, France
[3]Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia
[4]Software Science Department, Tallinn University of Technology, 12616 Tallinn, Estonia

Corresponding author: Rahul Sharma (rahul.sharma@taltech.ee)

This work has been partially conducted in the project ''ICT programme'' (information and communications technology programme) which was supported by the European Union through the European Social Fund.

ABSTRACT Statistical reasoning was one of the earliest methods to draw insights from data. However, over the last three decades, association rule mining and online analytical processing have gained massive ground in practice and theory. Logically, both association rule mining and online analytical processing have some common objectives, but they have been introduced with their own set of mathematical formalizations and have developed their specific terminologies. Therefore, it is difficult to reuse results from one domain in another. Furthermore, it is not easy to unlock the potential of statistical results in their application scenarios. The target of this paper is to bridge the artificial gaps between association rule mining, online analytical processing and statistical reasoning. We first provide an elaboration of the semantic correspondences between their foundations, i.e., itemset apparatus, relational algebra and probability theory. Subsequently, we propose a novel framework for the unification of association rule mining, online analytical processing and statistical reasoning. Additionally, an instance of the proposed framework is developed by implementing a sample decision support tool. The tool is compared with a state-of-the-art decision support tool and evaluated by a series of experiments using two real data sets and one synthetic data set. The results of the tool validate the framework for the unified usage of association rule mining, online analytical processing, and statistical reasoning. The tool clarifies in how far the operations of association rule mining and online analytical processing can complement each other in understanding data, data visualization and decision making.

INDEX TERMS Association rule mining, data mining, online analytical processing, statistical reasoning.

I. INTRODUCTION

Decision support techniques play an essential role in today's business environment. Since the 17th century, statistical reasoning (SR) has been used extensively to shape business decisions [1] and it was the earliest method to draw insights from data. With the emergence of decision support systems (DSSs) in the 1970s [2], SR is frequently used in DSSs and decision support tools, just take SPSS (Statistical Package for the Social Sciences) [3] or SAS (Statistical Analysis System) [4] as examples. With the rise of information technology in the 1990s, online analytical processing (OLAP) [5] and association rule mining (ARM) [6] have emerged as powerful decision support techniques (DSTs) [7], both with their specific rationales, objectives, and attitudes. Over the years, both OLAP and ARM have gained massive ground in practice (Cognos, SAP-BW resp. RapidMiner, Orange – to name a few) and, similarly, massive attention in the research community. Unfortunately, both OLAP and ARM have been introduced together with their own genuine mathematical formalizations and developed their specific terminologies. This makes it hard to reuse results from one domain in another; in particular, it is not always easy to unlock the potential of statistical results in OLAP and ARM application scenarios. OLAP represents relational data [8] in multi-dimensional views using roll-ups, drill-downs, slices, dices, etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

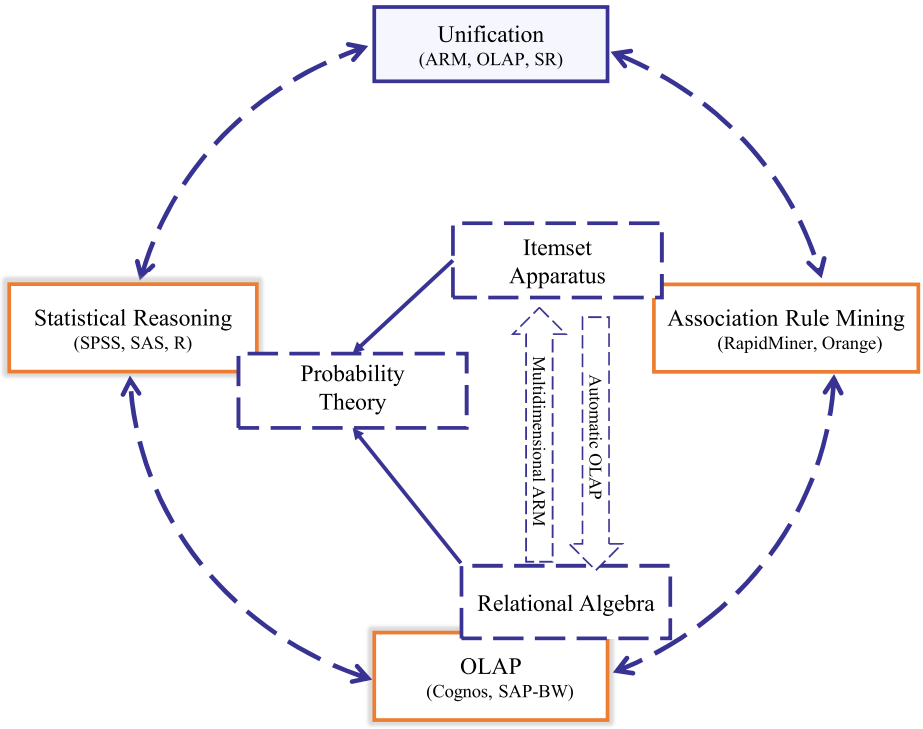


FIGURE 1. Semantic correspondences between association rule mining, online analytical processing and statistical reasoning.

In contrast, ARM relies on the notion of itemsets and frequent itemsets [9] in transaction databases. The correspondences between OLAP and ARM might seem rather simple, but it is neither fully elaborated in the state of the art nor implemented in practice. Because of the strong involvement of SR, OLAP, and ARM in decision-making, this paper aims to bridge the artificial gaps between them. We contribute by elaborating the semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., probability theory, relational algebra, and the itemset apparatus.

In Fig. 1, a graphical representation of the process of determining the semantic correspondence between the SR, OLAP, and ARM is shown. The solid rectangles are used to indicate the selected DSTs, and the blue dashed lines rectangles are used to indicate the foundations of DSTs. The adoption of concepts in between OLAP and ARM (and vice versa) is referred to as automatic OLAP [10] and multi-dimensional ARM [11], respectively. In Table 1 and Table 2, we provide a list of abbreviations and frequently used symbols that are being used throughout the paper.

In the process of establishing semantic correspondences between the three DSTs, probability theory and, in particular, conditional expected values (CEVs) are at the center of our considerations. CEVs correspond to *sliced average aggregates* in OLAP and would correspond to potential *ratio-scale confidences* in a generalized ARM [12]. Based on the semantic correspondences between the DSTs, we are convinced that it is possible to design advantageous next-generation features of advanced decision support tools. A series of popular decision support tools is given in Fig. 2. We use software polls by KDnuggets [13] in the years 2017, 2018, and 2019 to measure the popularity of these tools. The popularity percentages of the tools demonstrate that a diverse range of tools is popular in practice and that they have also gained massive attention in the research community.

Kamber *et al.* [11] addressed the integration of OLAP and ARM as soon as 1997. They have provided the notion of metarule-guided mining, which entails utilizing user-defined rule templates to direct the mining process. Later, Han *et al.* [14] have proposed DBMiner for interactive mining, which provides a wide range of data mining operations such as association, generalization, characterization, classification, and prediction. We also identify several approaches for integrating different DSTs, and there is significant research specifically on the integration of OLAP and ARM in state-of-the-art. We appraise all of these decision support frameworks and different ways of integrating DSTs; however, the concept of semantic correspondences between DSTs is yet to be elaborated in state-of-the-art. A detailed discussion on a variety of decision support frameworks and various approaches for the integration of DSTs is given in Sect. II. Elaborating the semantic correspondences between DSTs will be helpful to fill the artificial gaps between DSTs. Furthermore, it can enable decision-makers to work with cross-platform decision support tools and check their results from different viewpoints.

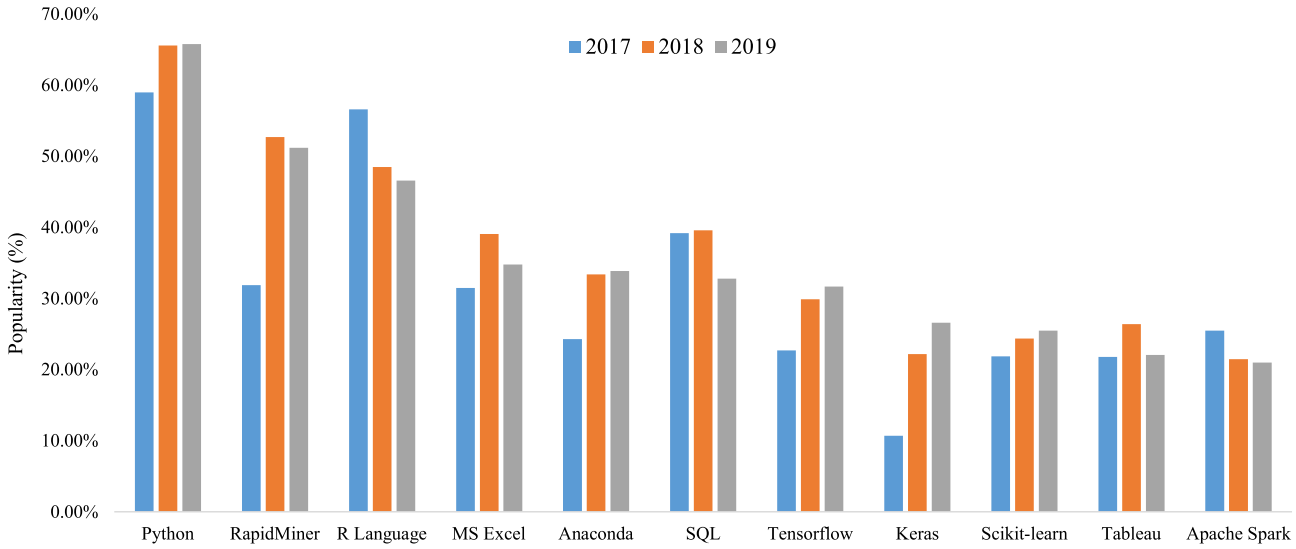


FIGURE 2. A series of popular decision support tools, together with their polarities according to opinion polls by KDnuggets [13] in 2017, 2018, and 2019.

TABLE 1. Abbreviations and acronyms.

Abbreviations	Summary
ACIF	All combination Influencing Factor
ARM	Association Rule Mining
CEVs	Conditional Expected Values
CFQs	Constrained Frequent Set
CAP	Consistency, Availability and Partition Tolerance
CBR	Case Based Reasoning
DIRECT	Discovering and Reconciling Conflicts
DMDSS	Data Mining Decision Support System
DST	Decision Support Technique
DSS	Decision Support System
FIA	Freedom of Information Act
GBP	Grid Based Pruning
GUI	Graphical User Interface
IADSS	Intelligent Agent Assisted DSS
IDSS	Integrated Decision Support System
KDD	Knowledge Discovery in Databases
MSMiner	Multi-strategy Data Mining Platform
NJ	New Jersey
OLAP	Online Analytical Processing
PVM	Parallel Virtual Machine
SAS	Statistical Software Suite
SAP	System Applications and Products in Data Processing
SAP-BW	SAP Business Warehouse
SPSS	Statistical Package for the Social Sciences
SR	Statistical Reasoning
uARMSolver	universal Association Rule Mining Solver
UDS1	User Defined Dataset

TABLE 2. List of frequent symbols.

Notation	Summary
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{B}$	Boolean values
$\mathbb{D}$	Discrete values
P	Probability
Σ	Events
Ω	Set of outcomes
T	Set of Transaction
D	OLAP Cube Dimension
σ	Relational Algebra
Δ	OLAP cube domain

The research on elaborating semantic correspondences between the three DSTs is significant due to the following reasons:

1) DSTs are developed independently for intended user groups and intended use cases.
2) Specific terminologies and functions of DSTs create artificial gaps between them and their tools.
3) Interpretation of results from one DST domain to another is not easily possible.
4) Artificial gaps between DSTs force decision-makers to use a variety of DSTs and decision support tools.
5) Various approaches for integrating DSTs are discussed in state of the art; however, correspondences between DSTs are obfuscated.

We observed that elaborating semantic correspondence between DSTs is necessary to bridge various artificial gaps between them. Therefore, in this paper, we elaborate semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., between probability theory, relational algebra, and the itemset apparatus. In particular, we formally establish the correspondence between (i) the support of an itemset and the probability of a corresponding event and (ii) the confidence of an association rule and the conditional probability of two corresponding events. And (iii), the OLAP average aggregate function turns out to correspond to conditional expected values, which closes the loop between ARM, OLAP, and probability theory with respect to the most important constructs in ARM and OLAP.

Based on the semantic correspondences between the DSTs, we propose a novel framework for the unification of DSTs. The framework provides a way to develop various next-generation decision support tools. To validate the proposed framework, we implement a sample tool by combining the operations of three DSTs. The tool's outcomes establish semantic correspondence between SR, OLAP, and ARM and provide various useful data visualization methods. The tool is implemented on ASP.NET. In the tool, we use *'all combinations of influencing factors'* (ACIF) function to select the target column and influencing factors to generate all possible combinations of data items. The programming code and other instructions on how to use the proposed tool are available in the GitHub repository [15]. We have named the tool *grand report* [12], [16]; a *grand report* provides a complete printout of generalized association rules, which can also be seen as the entire unfolding of a pivot table [17]. An instance of the tool is hosted and available on the web.[1] The tool is straightforward to use, and it provides unified usages of DSTs.

The key contributions of the paper are as follows:

1) Elaboration of semantic correspondences between the three DSTs, i.e., SR, OLAP, ARM, and their foundations, i.e., probability theory, relational algebra, and the itemset apparatus, respectively.
2) We characterize to what extent and how far SR, OLAP, and ARM can be considered synonymous.
3) A novel framework for the unification of DSTs is presented to develop next-generation decision support tools.
4) A sample tool is presented to implement the unification of DSTs. The tool provides unified usages of DSTs.
5) The tool is tested on various datasets and compared to a state-of-the-art decision support tool. The comparison and the tool's outcome demonstrate the tool's superior performance.

The paper is organized as follows: In Sect. II, we review current work related to the unification of SR, OLAP, and ARM. Then, in Sect. III, we discuss the main concepts of mainstream SR, OLAP, and ARM. In Sect. IV, we elaborate semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., probability theory, relational algebra, and the itemset apparatus. Subsequently, in Sect. V, we provide the framework for the unification of SR, OLAP, and ARM. A description of its implementation and experiments to showcase the relevance of the proposed framework are given. Finally, a discussion on future work and a conclusion are provided in Sect. VI and Sect. VII, respectively.

II. EXISTING WORK

In this section, previous work related to semantic correspondences between DSTs and various approaches for the integration of DSTs is explored.

[1] http://grandreport.me

The classical DSSs [2] were developed to assist managerial decisions by presenting several combinations of information. With the emergence of OLAP [5], knowledge discovery in databases (KDD) [36] and ARM [6], [37], many authors have proposed a variety of advanced DSSs. In the 1990s, web-based DSSs have been very popular [38]. Later, organizations have started taking advantage of different DSTs in DSSs [19]. We examine eighteen different research articles that discuss the integration of DSSs with different DSTs. A summary of these articles is given in Table 3. Wang [18] presented a novel architecture to integrate KDD techniques into existing DSSs. The authors have discussed the integration of different KDD techniques in group DSSs via three different types of decision support agents. In 2002, Fan *et al.* [19] provided a simple classification scheme for data value conflicts and presented an approach for discovering data conversion rules from data automatically. Bolloju *et al.* [20] provided a method for combining decision support and knowledge management to present an integrative framework for developing enterprise decision support environments. They used *model mart* and *model warehouse* as repositories.

In 2007, Rupnik *et al.* [24] discussed a method for combining DSS and data mining methods. The authors developed a data mining decision support system (DMDSS) that incorporates classification, clustering, and association rules. To investigate the use of data mining technology in DSS, Charest *et al.* [28] presented a theoretical, conceptual, and technological framework for the development of an intelligent data mining assistant by employing case-based reasoning and formal DL-ontology paradigms. Zhuang *et al.* [29] proposed a novel methodology to integrate data mining and case-based reasoning to develop a pathology test ordering system. In this paper, data mining concepts were used to extract the knowledge from past data, and then it was used in decision support.

In 2010, Liu *et al.* [30] conducted a survey to determine the efforts being made to develop an integrated decision support system (IDSS). IDSS combines four DSTs: knowledge-based systems, data mining, intelligent agents, and web technology. IDSS assists users in interpreting decision alternatives, and it also discovers hidden interesting patterns in large amounts of data using data mining tools. Gandhi *et al.* [39] demonstrated a DSS architecture (DSSA) that combines various data mining techniques. In this architecture, data mining tools were used to identify a set of features and patterns that domain experts can use to make decisions.

The majority of these works are inclined towards developing new DSSs and integrating DSSs with DSTs. However, the concept of semantic correspondences between DSTs is not discussed in any of these works. Therefore, we also explore the state of the art for the integration of OLAP and ARM. Some of these works focus on intra-dimensional association rules, while others are concerned with inter-dimensional association rules. Almost all intra-dimensional approaches use repeated predicates from a single data dimension.

TABLE 3. Existing approaches for the integration of different decision support techniques in DSSs.

Study	Year	Approach	Summary
Wang et al. [18]	1997	Intelligent agent-assisted DSS (IADSS)	A novel architecture is presented to integrate different KDD techniques in classical DSSs.
Fan et al. [19]	2002	A method for mining data value conversion rules from various data sources	In the process of integrating business data from multiple sources, a system called Discovering and Reconciling Conflicts (DIRECT) was presented .
Bolloju et al. [20]	2002	For the next generation DSSs, authors presented Integrating knowledge management into enterprise environments.	The authors proposed a method for combining decision support and knowledge management for developing enterprise decision support environments.
Heinrichs et al. [21]	2003	Integrating web-based data mining tools with DSS for knowledge management	Authors have highlighted how the knowledge workers in organizations can integrate data mining tools into their information and knowledge management requirements.
Cho et al. [22]	2003	Data mining for selection of insurance sales agents	For insurance managers, an intelligent DSS, Intelligent Agent Selection Assistant for Insurance, was presented.
Jukić et al. [23]	2006	Exploration of a large data warehouse using qualified association-rule mining	Mine fact tables and captures the correlations in data within data warehouses.
Rupnik et al. [24]	2007	DMDSS: a data mining-based DSS that combines data mining and decision support.	To support decision support procedures, a data mining-based DSS was provided.
March et al. [25]	2007	Integrated DSS: A data warehousing perspective	The authors have examined how data warehouses may be used for integration, implementation, intelligence, and innovation.
Shi et al. [26]	2007	MSMiner	The authors have presented an OLAP platform based on a data warehouse and integrated it with different data mining algorithms.
Domenica et al. [27]	2007	Stochastic programming and scenario generation within a simulation framework: An information systems perspective	This study discussed how sophisticated models and software realizations might be integrated with information systems and DSS tools.
Charest et al. [28]	2008	Intelligent data mining assistant using case-based reasoning	To study effectively deploying DM technology, the authors showed a framework for an intelligent data mining assistant by using case-based reasoning and formal DL ontology paradigms.
Zhuang et al. [29]	2009	Data mining and case-based reasoning for intelligent decision support	The authors used data mining and CBR approaches to provide information from previous data.
Liu et al. [30]	2010	IDSS	By utilizing data mining methods, IDSS assists users in interpreting decision alternatives and discovering hidden patterns in massive amounts of data.
Peng et al. [31],	2011	Incident information management framework	The authors presented a three level framework with three modules: data mining module, multi-criteria-decision-making module and data integration module.
Ltifi et al. [32].	2013	KDD-based human-centered design strategy for designing DSS	Based on KDD and other human-centered design principles, a human-centered design strategy for designing dynamic DSS was proposed
Dong et al. [33]	2014	A framework of Web-based decision support systems for portfolio selection with OLAP and PVM	The authors concentrated on developing a framework for a Web-based DSS.
Fister et al. [34]	2020	uARMSolver a framework	The paper introduces uARMSolver, a new software framework for ARM.
Aidan et al. [35]	2021	Knowledge Graph	Authors demonstrate how knowledge can be represented and extracted using a combination of deductive and inductive procedures.

A summary of different OLAP and ARM integration approaches is given in Table 4.

In 1998, Ng *et al.* [41], and Zhu [10] have proposed different ways to integrate ARM and OLAP together; however, their research was centered towards multi-dimensional ARM, automatic OLAP, and other specific sets of problems. The mainstream ARM was developed to find frequent items, while OLAP represented a multi-dimensional view of data using different OLAP operations. Therefore, the popularity of ARM for transactional datasets and the progress of OLAP [44] in a multi-dimensional environment attracted many authors to propose possible ways to integrate the ARM and OLAP. In 1997, Kamber *et al.* [11] first addressed the relationship between ARM and OLAP and proposed a meta-rule-guided mining approach for mining association rules from a multi-dimensional data cube. In this paper, Kamber *et al.* [11] have presented four algorithms that explore an OLAP data cube for meta-rule-guided mining of multi-dimensional association rules. Imielinski *et al.* [40] have presented cubegrades, a generalization of association rules which display how a set of measures (aggregates) is affected by specializing (rolldown), generalizing (roll-up) and mutating (which is a change in the cube's dimensions). In this paper, cubegrades are shown as more expressive than association rules in capturing associations and trends.

To support the adhoc mining in association rules, Lakshmanan *et al.* [42] proposed an idea of constrained frequent set queries (CFQs) and extended the architecture proposed by Ng *et al.* [41]. In addition, they introduced a new notion of quasi-succinctness and developed a heuristic technique for non-quasi-succinct constraints. Ng *et al.* [41] proposed architecture for exploratory mining of association

TABLE 4. Integration of OLAP and ARM in data mining.

Integration of ARM with OLAP	Approach	Summary
Kamber et al. [11]	Metarule-guided mining approach	Authors started by looking at the relationship between ARM and OLAP and then proposed a meta-rule-guided mining approach for extracting association rules from a multi-dimensional data cube.
Han et al. [14]	DBMiner: a software for various data mining techniques	DBMiner provides a way to combine data mining techniques with database technologies to uncover knowledge at different levels.
Imielinski et al. [40]	Cubegrades	To mine multi-dimensional association rules, a new concept of cubegrades with a novel grid-based pruning (GBP) technique was proposed as a generalized ARM technique. cubegrades was also shown as more expressive than associations rules in capturing associations and trends.
Raymond et al. [41]	Multidimensional Architecture, CAP algorithm, antimonocity, and succinctness	An architecture was proposed for multi-dimensional data mining, and CAP algorithm was used for the maximum degree of pruning. Two new rule pruning properties, antimonocity, and succinctness, were proposed to push the constraints deep inside the mining process.
Lakshmanan et al. [42]	Constrained Frequent Set (CFQs), the notion of quasi-succinctness, a heuristic technique for non-quasi-succinct constraints.	An idea of constrained frequent set queries (CFQs) was proposed In addition, they introduced a new notion of quasi-succinctness and developed a heuristic technique for non-quasi-succinct constraints.
Zhu et al. [10]	Multi-dimensional and Multi-level ARM methods	Proposed online analytical mining of association rules using the concept of multi-dimensional and multi-level ARM. Here, associations are divided into intra-dimensional, inter-dimensional, and constrained-based associations.
Nguyen et al. [43]	Exclusive Confidence and Support was proposed with a new algorithm	The authors proposed an architecture that allows constraint-based and human-centered exploratory mining of association rules. Two new measures, exclusive confidence and natural confidence, were discussed.

rules that is constraint-based and human-centered. To push the constraints deep inside the mining process, this paper presents a new algorithm (CAP) and two new rule pruning properties; antimonocity and succinctness. To generalize ARM within arbitrary n-ary relations and boolean tensors, Nguyen *et al.* [43] proposed exclusive confidence and natural confidence measures. They have also designed a complete, scalable algorithm that computes the exclusive measures. Kamber *et al.* [11] extended the constrained gradient analysis ''cubegrades'' presented by Imielinski *et al.* [40]. In this paper, the authors have addressed various issues and methods on efficient mining of multi-dimensional, constrained gradients in multi-dimensional data cubes. They have also defined the constraints as significant constraints, probe constraints, and gradient constraints.

Zhu [10] proposed online analytical mining of association rules and presented a step-by-step method and algorithm for inter-dimensional ARM, intra-dimensional ARM, and hybrid ARM. Based on OLAP technologies, they also designed a method to perform multi-level ARM. Chen *et al.* [45] developed an OLAP and data warehousing-based platform for weblog records (WLRs), which supports multi-level and multi-dimensional ARM. Finally, Cerf *et al.* [46] have presented an n-array algorithm for n-array relations, which was used to extract constrained-based closed n-sets.

In the state of the art, integration of DSTs and DSSs frameworks are broadly discussed. However, the correspondences between the foundation of DSTs are obfuscated. Therefore, we aim to elaborate semantic correspondences between the foundations of the three popular DSTs and bridge the artificial gaps between them.

III. PRELIMINARIES

This section provides background information about the three popular DSTs, i.e., SR, OLAP, ARM and their foundation, i.e., probability theory, relational algebra, and itemset mining. In Sect. III-A, we discuss the concepts of SR. Then, in Sect. III-B, the concepts of classical ARM are discussed, and in Sect. III-C, we discuss the basic concepts of OLAP.

A. STATISTICAL REASONING (SR)

With the development of probability theory [1] by thinkers like Gerolamo Cardano, Blaise Pascal, and Pierre de Fermat, statistics has evolved as an essential framework for developing DSS [47] and DSTs; therefore, most of the DSTs have been developed with the core concepts of SR. Since 1970, extensive use of computer systems has made it possible to do large statistical computations that have not been possible manually. In the 19^{th} and 20^{th} centuries, statistics had its victory by evolving into the primary scientific tool – think about classical thermodynamics and its elaboration through statistical mechanics and quantum physics. In the natural sciences, statistics have become the necessary foundation in economics, and many Nobel prizes correspond with the probabilistic variants of game theory. So, it could be said that statistics is the language of science. However, even more, statistics was a crucial driver in the industrial revolution, by helping to optimize production, think about Student's t-distribution.

Moreover, statistics is at the core of optimizing production; think of Six Sigma alone. All this is true, but since 1970, we have seen the next wave of SR. Statistics has left the scientific laboratories and entered the everyday decision-making

TABLE 5. Types of input data used in various decision support techniques.

Decision Support techniques	Types of Input data
Statistical Reasoning: Classification	$X_1 : T_1 \times \ldots \times X_m : T_m$
Statistical Reasoning: Multivariate Data Analysis (Regression)	$X_1 : \mathbb{R} \times \ldots \times X_m : \mathbb{R}$
Online Analytical Processing (OLAP)	$\underbrace{X_1 : D_1 \times \ldots \times X_m : D_m}_{\text{discrete}} \times \underbrace{Y_1 : \mathbb{R} \times \ldots \times Y_{m'} : \mathbb{R}}_{\text{numerical}}$
Association Rule Mining (Bitmap)	$X_1 : \mathbb{B} \times \ldots \times X_m : \mathbb{B}$
Association Rule Mining in Tools	$\underbrace{X_{1_1} : \mathbb{B} \times \ldots \times X_{n_1} : \mathbb{B}}_{X_1 : D_1} \times \cdots \times \underbrace{X_{m_1} : \mathbb{B} \times \ldots \times X_{n_m} : \mathbb{B}}_{X_m : D_m}$

processes in our organizations. Here, SR is the tool of highly specialized experts in highly specialized tasks but becomes available to a broader range of decision-makers. This movement is precisely about what has been expressed by ''The Future of Data Analysis'' by Tukey [48]. It means that systematic decision-making becomes more and more pervasive. In our opinion, this also explains the emergence of ARM and OLAP, which are two immensely successful approaches that complement, extend (but also overlap) the established SR toolkit. Moreover, the journey has just begun, as the current interest in *data science* proves – in 2015, Donoho [49] showed the evolution of data science from statistics. In Table 5, we provide different combinations of data used in SR, OLAP, and ARM. $\mathbb{R}$ is used to represent numerical type data, $\mathbb{D}$ is used to represent discrete type data and $\mathbb{B}$ is used to represent bitmap data.

B. ASSOCIATION RULE MINING (ARM)

To understand the relationship between different data items in transactional datasets and to find out interesting patterns and correlations, Agrawal *et al.* [6] presented the central concept of ARM using binary representations of data items as shown in Table 5. However, ARM is also presented for numerical data items as quantitative ARM [50], numerical ARM [51], [52].

ARM is highly effective in discovering relations and interesting associations among data items using different measures of interestingness [6], [53] and it is a prevalent technique that plays a crucial role in market basket data analysis, bioinformatics, ocean, land, and medical diagnosis.

In the original settings, association rules are extracted from transactional datasets composed of a set $I = \{i_1, \ldots, i_n\}$ of n binary attributes called *items* and a set $D = \{t_1, \ldots, t_n\}$, $t_k \subseteq I$, of *transactions* called database. An *association rule* is a pair of itemsets (X, Y), often denoted by an implication of the form $X \Rightarrow Y$, where X is called the antecedent (or premise) and Y is called the consequent (or conclusion), $X \cap Y = \emptyset$. To select interesting association rules, the following are the most popular measures of interestingness in ARM.

Definition 1: The *Support* of an itemset X with respect to a set of transactions T, denoted by $Supp(X)$, is the ratio of transactions that contain all items of X (number of transactions that satisfy X) [54]:

$$\text{Supp}(X) = \frac{|\{t \in T \mid X \subseteq t\}|}{|T|}$$

Definition 2: The *confidence* of an association rule $X \Rightarrow Y$ concerning a set of transaction T, denoted by $Conf(X \Rightarrow Y)$ is the percentage of transactions that contains X which also includes Y. Technically, the confidence of an AR is an estimation of the conditional probability of Y over X:

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}.$$

Definition 3: The *lift* of an association rule $X \Rightarrow Y$, denoted by $Lift(X \Rightarrow Y)$, is used to measure misleading rules that satisfy minimum support and minimum confidence threshold. The Lift measure is also used to calculate the deviation between an antecedent X and a consequent Y, which is the ratio of the joint probability of X and Y divided by the product of their marginal probabilities.

$$Lift(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X) \times Supp(Y)}$$

In ARM, when the number of association rules is too large to be presented to a data mining expert or even treated by a computer, measures of interestingness can filter the interesting association rules. After support, confidence, and lift, more than fifty different measures of interestingness are in the literature [53], [55], [56]. These measures of interestingness are discussed in detail in the literature [57], [58]. Initially, ARM was limited to large transactional datasets. Still, later, Han *et al.*, Lu *et al.*, Imielinski *et al.*, and Nguyen *et al.* [40], [43], [59], [60] presented different views on multi-level and multi-dimensional ARM. Over the years, different ARM frameworks [34] and the use of ARM in varied application scenarios [61], [62] have also been discussed in the state of the art [63].

1) MULTIDIMENSIONAL VIEW OF ARM

More recently, ARM has been adapted to the multidimensional case [43] and multitask-based ARM [64]. In multidimensional setting of ARM, datasets are composed of a set $\mathcal{D} = \{S_1, \ldots, S_n\}$ of *dimensions*, and an n-ary relation between them, i.e. they are formally tuples $(S_1, \ldots, S_n, R)$ with $R \subseteq S_1 \times \cdots \times S_n$. In "Multitask-based" ARM, highly frequent association rules for different ARM tasks are referred as "single-task" rules which are Later combined together to generate the global results, i.e, "multitask rules".

Multidimensional association rules are rules between two so-called *associations* that generalize the notion of itemset. They are defined as the Cartesian products of subsets of dimensions. The set of dimensions used in an association X is called its *domain* and is noted $dom(X)$. For example, $X = \{Milk, Bread\} \times \{Winter\}$ is an association on the domain $dom(X) = \{products, seasons\}$. We use $\pi_{S_i}(X)$ to denote the projection of the association X on the dimension S_i, e.g. $\pi_{products}(X) =$ {Milk, Bread} and $\pi_{seasons}(X) =$ {Winter}.

In the multi-dimensional case, the generalization of the notion of support is the *relative support*. The *support of an association X relative to a set D* $\supseteq dom(X)$ *of dimensions* is defined as

$$Supp_D(X) = \Big|\{t \in \prod_{S_d \in \mathcal{D}\setminus\{D\}} S_d \quad | \quad \exists u \in \prod_{S_i \in D\setminus dom(X)} S_i \text{ such that } \forall x \in X, x.u.t \in \mathcal{R}\}\Big| \quad (1)$$

Using the relative support, two variants of confidence, the *exclusive confidence* and *natural confidence* are defined for multidimensional association rules:

$$Conf_{natural}(X \Rightarrow Y) = \frac{Supp_{dom(X\cup Y)}(X \cup Y)}{Supp_{dom(X\cup Y)}(X)}$$

$$Conf_{exclusive}(X \Rightarrow Y) = \frac{Supp_{dom(X\cup Y)}(X \cup Y) \times P}{Supp_{dom(X)}(X)}$$

with $P = |\prod_{S_i \in dom(X\cup Y)\setminus dom(X)} \pi_{S_i}(Y)|$.

In Table 6, the multidimensional association rule $\{Milk\} \Rightarrow \{Bread\} \times \{Spring\}$ has a natural support of $\frac{1}{4}$ because

$$Supp_{\{products,seasons\}}(\{Milk, Bread\} \times \{Spring\}) = |\{c_2\}| = 1$$
$$Supp_{\{products,seasons\}}(\{Milk\}) = |\{c_1, c_2, c_3, c_4\}| = 4. \quad (2)$$

This rule can also be expressed in first-order logic, i.e.

$$\{Milk\} \Rightarrow \{Bread\} \times \{Spring\} \equiv \forall X, Y, \neg purchase(X, Milk, Y) \vee (purchase(X, Bread, Spring) \wedge purchase(X, Milk, Spring)). \quad (3)$$

C. ONLINE ANALYTICAL PROCESSING (OLAP)

Historically, OLAP is not a new idea; it has persisted over the decades. Initially, in 1962, Kenneth Iverson proposed the foundation of OLAP in his book "A Programming Language" [65]. In 1975, Information Resources Inc. launched the first OLAP product named "Express", which was acquired by Oracle Inc. in 1995. In 1993, Edgar F. Codd used the term OLAP and set up 12 policies for an OLAP product in his paper "Providing OLAP (Online Analytical Processing) to user-analysts: An IT mandate" [5]. In OLAP, it is essential to have a multi-dimensional cube. Therefore, we show a sample OLAP cube with three dimensions (D_1, D_2, D_3) in Fig. 3. Practically, an OLAP cube consists four types of functions; First, OLAP operations, i.e., RollUp, Drill Down, Slice, Dice, and Pivot. Second is aggregation operations, i.e., SUM, AVG, COUNT, MIN, MAX, calculate trends, ranking, percentiles, attribute-based grouping, compare aggregates, etc. The third is the OLAP operator, i.e., "Force" and "Extract," which convert a dimension into a measure and a measure into a dimension. Fourth is the capability to handle uncertain data within the OLAP model.

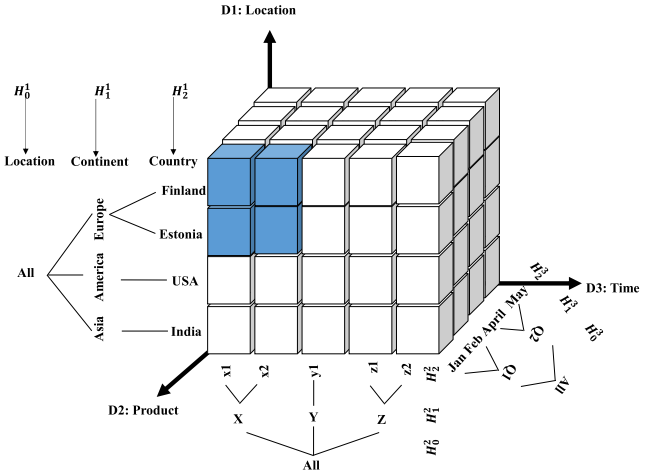


FIGURE 3. A sample OLAP data cube with three dimensions (D1: location, D2: product and D3: time).

IV. SEMANTIC CORRESPONDENCE BETWEEN SR, OLAP AND ARM

In this section, we establish semantic correspondence between SR, OLAP, and ARM. We use probability theory with conditional expected values (CEVs) as the center of our mappings. First, we provide semantic correspondence between SR, i.e., probability theory and ARM, and then we provide semantic correspondence between SR and OLAP.

Definition 4 (σ-Algebra): Given a set Ω, a σ*-Algebra* Σ *over* Ω is a set of subsets of Ω, i.e., $\Sigma \subseteq \mathsf{P}(\Omega)$, such that the following conditions hold true:

1) $\Omega \in \Sigma$
2) If $A \in \Sigma$ then $\Omega\setminus A \in \Sigma$
3) For all countable subsets of Σ, i.e., $A_0, A_1, A_2 \ldots \in \Sigma$ it holds true that $\bigcup_{i\in\mathbb{N}_0} A_i \in \Sigma$

Definition 5 (Probability Space): A *probability space* $(\Omega, \Sigma, \mathsf{P})$ consists of a set of outcomes Ω, σ algebra of (random) events Σ over the set of outcomes Ω and a probability function P: $\Sigma \to \mathbb{R}$, also called probability measure, such that the following axioms hold true:

TABLE 6. A multidimensional binary dataset in which *customers* (c_1 to c_4) buy *products* (Milk, Bread, Diapers, Beer) during *seasons* (Winter, Spring, Summer).

	Milk	Bread	Diapers	Beer	Milk	Bread	Diapers	Beer	Milk	Bread	Diapers	Beer
c_1	1	1	0	0	1	0	1	0	0	0	1	0
c_2	0	0	1	1	1	1	0	0	1	0	1	0
c_3	1	1	1	1	0	1	1	1	1	1	0	0
c_4	1	1	0	1	1	0	0	1	0	0	1	1
	Winter				Spring				Summer			

1) $\forall A \in \Sigma . 0 \leq \mathsf{P}(A) \leq 1$ (i.e., $\mathsf{P}\colon \Sigma \to [0, 1]$)
2) $\mathsf{P}(\Omega) = 1$
3) (Countable Additivity): For all countable sets of pairwise disjoint events, i.e., $A_0, A_1, A_2 \ldots \in \Sigma$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$, it holds true that

$$\mathsf{P}\Big(\bigcup_{i=0}^{\infty} A_i\Big) = \sum_{i=0}^{\infty} \mathsf{P}(A_i)$$

Definition 6 (Conditional Probability): Given two events $\{X, Y\} \in \Sigma$ of *probability space* $(\Omega, \Sigma, \mathsf{P})$. If $\mathsf{P}(X) \neq 0$ then we define conditional probability of Y given X as:

$$\mathsf{P}(Y|X) = \frac{\mathsf{P}(X \cap Y)}{\mathsf{P}(X)}$$

Definition 7 (Expected Value): Given a real-valued discrete random variable $X\colon \Omega \to I$ with indicator set $I = \{i_0, i_1, i_2, \ldots, i_n\} \subseteq \mathbb{R}$ based on $(\Omega, \Sigma, \mathsf{P})$, the *expected value* $\mathsf{E}(X)$, or *expectation* of X (where E can also be denoted as E_P in so-called explicit notation) is defined as follows:

$$\mathsf{E}(X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(X = i_n)$$

Definition 8 (Conditional Expected Value): Given a real-valued discrete random variable $Y\colon \Omega \to I$ with indicator set $I = \{i_0, i_1, i_2, \ldots\} \subseteq \mathbb{R}$ based on a probability space $(\Omega, \Sigma, \mathsf{P})$ and an event $X \in \Sigma$, the *expected value* $\mathsf{E}(Y)$ of Y conditional on X (where E can also be denoted as E_P in so-called explicit notation) is defined as follows:

$$\mathsf{E}(Y|X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(Y = i_n | X) \tag{4}$$

A. ANCHORING ASSOCIATION RULE MINING IN PROBABILITY THEORY

We follow the concepts and notation and their formalization as originally introduced by Agrawal *et al.* in their 1993 paper [6] as closely as possible. First, there is a *whole itemset* $\mathfrak{I} = \{I_1, I_2, \ldots, I_n\}$ consisting of a *total number n* of items $I_1, I_2, \ldots, I_n$. A subset $X \subseteq \mathfrak{I}$ of the whole itemset is called an *itemset*. Next, we introduce the notion of a *set of transactions T* (*that fits the itemset* $\mathfrak{I}$) as a relation as follows:

$$T \subseteq TID \times \underbrace{\{0, 1\} \times \cdots \times \{0, 1\}}_{n-\text{times}} \tag{5}$$

Here, *TID* is a finite set of transaction identifiers. For the sake of convenience, we assume that it has the form $TID = \{1, \ldots, N\}$. Actually, we need to impose a uniqueness constraint on *TID*, i.e., we require that T is right-unique, i.e., a function given as,

$$T \in TID \longrightarrow \underbrace{\{0, 1\} \times \cdots \times \{0, 1\}}_{n-\text{times}} \tag{6}$$

Given (6), we have that N in $TID = \{1, \ldots, N\}$ equals the size of T, i.e., $N = |T|$. Henceforth, we refer to T interchangeably both as a relation and as a function, according to (5) resp. (6). For example, we use $t = \langle i, i_1, \ldots i_n \rangle$ to denote an arbitrary transaction $t \in T$; similarly, we use $T(i)$ to denote the i-th transaction of T more explicitly etc. Given this formalization of the transaction set T, it is correct to say that T is a binary relation between TID and the whole itemset. In that, $I_1, I_2, \ldots, I_n$ need to be thought of as column labels, i.e., there is exactly one bitmap column for each of the n items in $\mathfrak{I}$, compare with (5) and (6). Similarly, Agrawal *et al.* have called the single transaction a bit vector and introduced the notation $t[k]$ for selecting the value of the transaction t in the k-th column of the bitmap table (in counting the columns of the bitmap table, the *TID* column is omitted, as it merely serves the purpose of providing transaction identities), i.e., given a transaction $\langle tid, i_1, \ldots i_n \rangle \in T$, we define $\langle tid, i_1, \ldots i_n \rangle[k] = i_k$. Less explicit, with the help of the usual tuple projection notation π_j, we can define $t[k] = \pi_{k+1}(t)$. Let us call a pair $\langle \mathfrak{I}, T \rangle$ of a whole itemset $\mathfrak{I}$ and a set of transaction T that fits $\mathfrak{I}$ as described above an *ARM frame*. Henceforth, we assume an ARM frame $\langle \mathfrak{I}, T \rangle$ as given.

We have said that a transaction is a bit vector. For the sake of convenience, let us introduce some notation that allows us to treat a transaction as an itemset. Given a transaction $t \in T$ we denote the *set of all items that occur in t* as $\{t\}$ and we define it as follows:

$$\{t\} = \{I_k \in \mathfrak{I} \mid t[k] = 1\} \tag{7}$$

The {t} notation provided by (7) will prove helpful later, as it allows us to express properties about transactions without the need to use bit-vector notation, i.e., without the need to maintain item numbers k of items I_k.

Given an $I_j \in \mathfrak{I}$ and a transaction $t \in T$, Agrawal *et al.* says [6] that I_j *is bought by* t if and only if $t[j] = 1$. Similarly, we can say that t *contains* I_j in such case. Next, given an itemset $X \subseteq \mathfrak{I}$ and a transaction $t \in T$, Agrawal *et al.* says

that t *satisfies* X if and only if $t[j] = 1$ for all $I_j \in X$. Similarly, we can say that t *contains all of the items of* X in such case. Next, we can see that t satisfies X if and only if $X \subseteq \{t\}$. Henceforth, we use $X \subseteq \{t\}$ to denote that t satisfies X.

Given an itemset $X \subseteq \mathfrak{I}$, the relative number of all transactions that satisfy X is called the *support of* X and is denoted as $Supp(X)$, i.e., we define:

$$Supp(X) = \frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{8}$$

Again, it makes perfect sense to talk about the support of an itemset X as the relative number of all transactions that each contain all of the items of X.

An ordered pair of itemsets $X \subseteq \mathfrak{I}$ and $Y \subseteq \mathfrak{I}$ is called an *association rule*, and is denoted by $X \Rightarrow Y$. Now, the relative number of all transactions that satisfy Y among all of those transactions that satisfy X is called the *confidence of* $X \Rightarrow Y$, and is denoted as $Conf(X \Rightarrow Y)$, i.e., we define:

$$Conf(X \Rightarrow Y) = \frac{|\{\, t \in T \mid Y \subseteq \{t\} \ \wedge\ X \subseteq \{t\}\,\}|}{|\{t \in T \mid X \subseteq \{t\}\}|} \tag{9}$$

Usually, the confidence of an association rule is introduced via support of itemsets as follows:

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \tag{10}$$

It can easily be checked that (9) and (10) are equivalent.

B. SEMANTIC CORRESPONDENCE BETWEEN ARM AND SR)

Next, we map the concepts defined in ARM to probability theory. Given an ARM frame $F = \langle \mathfrak{I}, T \rangle$ next we map the concepts defined in ARM to probability space $(\Omega_F, \Sigma_F, \mathsf{P}_F)$. First, we define the set of outcomes Ω_F to be the set of transactions T. Next, we define Σ_F to be the power set of Ω_F. Finally, given an event $X \in \Sigma_F$, we define the probability of X as the relative size of X, as follows:

$$\Omega_F = T \tag{11}$$

$$\Sigma_F = \mathbb{P}(T) \tag{12}$$

$$\mathsf{P}_F(X) = \frac{|X|}{|T|} \tag{13}$$

In the sequel, we drop the indices from Ω_F, Σ_F, and P_F, i.e., we simply use Ω, Σ, and P to denote them, but always keep in mind that we actually provide correspondence from ARM frames F to corresponding probability spaces $(\Omega_F, \Sigma_F, \mathsf{P}_F)$. The idea is simple. Each transaction is modeled as an outcome and, as usual, also a basic event. Furthermore, each set of transactions is an event.

We step forward with item and itemsets. For each item $I \in \mathfrak{I}$ we introduce the *event that item* I *is contained in a transaction*, and we denote that event as $[\![I]\!]$. Next, for each itemset $X \subseteq \mathfrak{I}$, we introduce the *event that all of the items in* X *are contained in a transaction* and we denote that event as $[\![X]\!]$. We define:

$$[\![I]\!] = \{\, t \mid I \in \{t\}\,\} \tag{14}$$

$$[\![X]\!] = \bigcap_{I \in X} [\![I]\!] \tag{15}$$

As usual, we identify an event $[\![I]\!]$ with the characteristic random variable $[\![I]\!] : \Omega \longrightarrow \{0, 1\}$ and use $\mathsf{P}([\![I]\!])$ and $\mathsf{P}([\![I]\!] = 1)$ as interchangeable.

1) FORMAL CORRESPONDENCE OF ARM SUPPORT AND CONFIDENCE TO PROBABILITY THEORY

Based on the correspondence provided by (11) through (15), we can see how ARM *Support* and *Confidence* translate into probability theory.

Lemma 1 (Mapping ARM Support *to Probability Theory):* Given an itemset $X \subseteq \mathfrak{I}$, we have that:

$$Supp(X) = \mathsf{P}([\![X]\!]) \tag{16}$$

Proof: According to (15), we have that $\mathsf{P}([\![X]\!])$ equals

$$\mathsf{P}\Big(\bigcap_{I \in X} [\![I]\!]\Big) \tag{17}$$

Due to (14), we have that (17) equals

$$\mathsf{P}\Big(\bigcap_{I \in X} \{\, t \in T \mid I \in \{t\}\,\}\Big) \tag{18}$$

We have that (18) equals

$$\mathsf{P}(\{\, t \in T \mid \bigwedge_{I \in X} I \in \{t\}\,\}) \tag{19}$$

We have that (19) equals

$$\mathsf{P}(\{t \in T \mid X \subseteq \{t\}\}) \tag{20}$$

According to (13), we have that (20) equals

$$\frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{21}$$

According to (8), we have that (21) equals $Supp(X)$

Lemma 2 (Mapping ARM Confidence *to Probability Theory):* Given an itemset $X \subseteq \mathfrak{I}$, we have that:

$$Conf(X \Rightarrow Y) = \mathsf{P}\big(\, [\![Y]\!] \,\big|\, [\![X]\!] \,\big)$$

Proof: Omitted.

In Table 7, we provide one to one mapping in between the operations of ARM and SR, i.e., probability theory. A set of items in ARM $\mathfrak{I} = \{I_1, I_2, \ldots, I_m\}$ are equivalent to the set of events $\mathfrak{I} = \{I_1 \subseteq \Omega, \ldots, I_m \subseteq \Omega\}$ in probability theory. Transactions T in ARM are equivalent to the set of outcomes Ω in probability space $(\Omega, \Sigma, \mathsf{P})$. Support of an itemset X in ARM is equivalent to the relative probability of the itemset X. Confidence of an association rule $X \Rightarrow Y$ is equivalent to the conditional probability of Y in the presence of X.

C. ANCHORING OLAP IN PROBABILITY THEORY

Decision-makers are using OLAP to explore data in a multi-dimensional view. It helps to compute different aggregate summaries using various OLAP operations (COUNT, SUM, Drill-Down, Roll-up, Slice, Dice, etc.). For example, Fig. 4 demonstrates age and salary records in a two-dimensional space. In OLAP, data exploration starts from a high granularity level to a lower granularity level or vice versa. The sample data cube is given in Fig. 3 consists of *time*, *location* and *product* dimensions. An OLAP dimension comprises

TABLE 7. Semantic correspondences between association rule mining and statistical reasoning (probability theory).

Association Rule Mining Terminology	Statistical Reasoning (Probability Theory)
Set of Items $\mathfrak{I} = \{I_1, I_2, \ldots, I_m\}$	Set of Events $\mathfrak{I} = \{I_1 \subseteq \Omega, \ldots, I_m \subseteq \Omega\}$
Transactions $T \subseteq I_0 \times \underbrace{\{0,1\} \times \ldots \times \{0,1\}}_{m-times}$	Ω
$t \in T$ *satisfies* itemset $X \subseteq \mathfrak{I}$: $t[k] = 1 \Rightarrow I_k \in X$	$t \in \cap X$
Support of itemset $X \subseteq \mathfrak{I}$: $\text{Supp}(X) = \frac{\lvert\{t \in T \mid X \subseteq t\}\rvert}{\lvert T\rvert}$	$\mathsf{P}(\bigcap_{I \in X} [\![I]\!])$
Confidence of association rule $X \Rightarrow Y$: $\text{Conf}(X) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$ $X \subseteq T, Y \subseteq T$, usually: $X \cap Y = \emptyset, \lvert Y\rvert = 1$	Conditional Probability $\mathsf{P}([\![Y]\!] \mid [\![X]\!])$
Lift of $X \Rightarrow Y$ wrt the outermost margin: $\frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$	$\frac{\mathsf{P}(\cap Y \mid \cap X)}{\mathsf{P}(\cap Y)}$

organized attributes in a hierarchical structure to show the different data granularity levels. For example, the time dimension in Fig. 3 may have the following hierarchy: *Month* $\rightarrow$ *Quarter* $\rightarrow$ *Year*. Here, the dimension attribute *Year* shows a high level of granularity, and *Month* shows a lower level of granularity. Based on the sample OLAP cube given in Fig. 3, first, we provide standard notions and definitions and then provide semantic correspondence between OLAP and SR, i.e., probability theory.

1) OLAP CUBE: BASIC NOTATIONS AND DEFINITIONS

Let an OLAP cube C be a multi-dimensional data cube with four-tuple $C = \{\Delta, D, H, M\}$ where Δ represents the OLAP cube domain, D is a non-empty set of n dimensions, H is a set of dimension hierarchy and M is a non-empty set of quantitative measures, i.e., numerical or additive values of a cell. We have considered the following properties concerning the OLAP cube.

- In an OLAP cube C, dimension set $D = \{D_1, D_2 \ldots D_i \ldots D_n\}$, dimension D_i consists of a set of different hierarchy levels H_i, where $i \leq n$.
- A hierarchy level $H_j^i \in H_i$ is a non-empty set of members A_{ij}. $H_j^i (j \geq 0)$ is the j^{th} hierarchical level in D_i. E.g., in Fig. 3, the set of hierarchical level of dimension D_1 is $H_1 = \{H_0^1, H_1^1, H_2^1\} = \{Location, Continent, Country\}$, and in the dimension D_1, the set of members at level H_2^1 is $A_{12} = \{India, USA, Estonia, Finland\}$

Definition 9: Sub Cube: A sub cube C' is part of the main OLAP cube with a non-empty set D' of m dimensions. $D' = \{D_1, D_2 \ldots D_i \ldots D_m\}$ and $m \leq n$. According to D', a tuple $\{\Theta_1 \ldots \Theta_m\}$ is a sub cube C' if $D' \subseteq D$ and $\Theta_1 \subseteq A_{ij}$ for all $i \in \{1 \ldots m\}$ and $\Theta_i \neq null$.

E.g., If in Fig. 3, a dimension set $D' = \{D1, D2\} \in D$ is a sub cube then $(\Theta_1, \Theta_2) = \{Europe, x_1, x_2\}$ will be a sub cube.

Definition 10: Aggregate Measure: A Measure M in a data cube C is the SUM of measure M of all facts in the cube.

E.g., "Total Sales" in Fig. 3 can be evaluated by its sum-based aggregate measure. The aggregate expression

Age	Salary
18	10000
20	12000
25	20000
30	25000
35	35000
40	40000
45	50000
50	70000
55	90000
60	120000

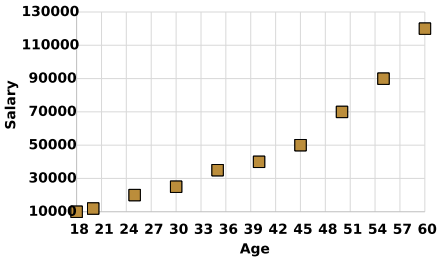


FIGURE 4. A sample representation of age and salary records in two dimensional space.

TotalSales(*India*, $\{x_1, x_2, y_1\}$) represents the SUM of total sales turnover for the products (x_1, x_2, y_1) in India.

Definition 11: Intra Dimension Predicate: A dimension predicate A_i in a dimension D_i is its member as a value represented as $a_i \in A_{i_j}$.

E.g., In Fig. 3, a dimension predicate a_1 in dimension D_1 is $a_1 \in \{Asia, America, Europe\}$.

Definition 12: Inter Dimension Predicate: Let data cube C have a sub cube C' with a non empty set of dimensions $D' = \{D_1, D_2 \ldots D_i \ldots D_m\}$ and $D' \subseteq D$. When the value of dimension predicates $\{A_1 \ldots A_m\}$ belongs to two or more dimensions where $(2 \leq m \leq n)$, then it is referred to as inter dimension predicates.

E.g., In Fig. 3, dimension predicate $\{a_1, a_2\} \in \{D_1, D_2\}$ then $a_1 \in \{Asia, America, Europe\}$ and $a_2 \in \{X, Y, Z\}$.

2) SEMANTIC CORRESPONDENCE BETWEEN OLAP AND SR

As discussed in Sect. III-C, an OLAP cube consists of various operations (Roll-Up, Drill-Down, Slice, Dice, Pivot, SUM, AVG, MIN, etc.). We have that the OLAP conditional operations (Slice, Dice, Drill- Down, Roll-up) on bitmap (Binary) columns correspond to conditional probabilities. Those conditional operations on numerical columns correspond to conditional expected values in probability theory. For example, we model a sample OLAP Table 8 in probability theory. We consider that Table 8 is equivalent to the set of outcomes.

TABLE 8. A sample OLAP table.

City	Profession	Education	Age Group	Freelancer	Salary
New York	Lawyer	Master	25–30	0	3, 800
Seattle	IT	Bachelor	18–25	1	4, 200
Boston	Lawyer	PhD	40–50	1	12, 700
L.A.	Chef	High School	30–40	0	3, 700
...	...	...	...	...	...

Ω in probability space (Ω, Σ, P), a row r is an element of Ω, i.e. $r \in \Omega$ and each column c is equivalent to a random variable $\mathbb{R}$. We consider numerical columns as *finite real-valued* random variables (For Example: Salary $\in \Omega \subseteq \mathbb{R}$) and bitmap columns are considered as events (For Example: Freelancer $\subseteq \Omega$). The following is a probabilistic interpretation of the OLAP Table 8.

- City: $\Omega \longrightarrow \{Boston, L.A., New\ York.\ldots\ldots\}$
- Profession: $\Omega \longrightarrow \{Chef, Construction.\ldots..\}$
- Education Level: $\Omega \longrightarrow \{High\ School.\ldots..\}$
- Age Group: $\Omega \longrightarrow \{18–20, 25–30\ldots > 65\}$
- Freelancer: $\Omega \longrightarrow \{0, 1\}$
- Salary: $\Omega \longrightarrow I_{\text{Salary}} \subseteq \mathbb{R}(|I_{\text{Salary}}| \in \mathbb{N})$

3) SEMANTIC CORRESPONDENCE BETWEEN OLAP AVERAGES AND SR

In many cases and as per Codd *et al.* [5], decision-makers use SQL queries to interact with OLAP. Therefore, we start with simple OLAP queries mapped with probability theory. We have a simple OLAP average query; (SELECT AVG(Salary) FROM Table 8). If the number of rows of Table 8 is represented by $|\Omega|$ and the number of rows that contain a value i in column C are equivalent to $\#_C(i)$ then AVG(Salary) FROM Table 8 will compute the average of all the salaries, i.e., a fraction of the sum of the column (Salary) and the total number of rows in the table.

In probability theory, the *average* of a random variable X is the *Expected Value* of $X = \mathsf{E}[X]$. We compare the expected value of X, i.e., $\mathsf{E}(X)$ with the output of the AVG query in OLAP. We have *OLAP Query:*

$$(SELECT\ AVG(Salary)\ FROM\ Table\ 8) \tag{22}$$

$$\text{Expected Value: } \mathsf{E}(\text{Salary}) = \sum_{i \in I_{\text{Salary}}} i \cdot \mathsf{P}(\text{Salary} = i) \tag{23}$$

$$= \sum_{i \in I_{\text{Salary}}} i \cdot \frac{\#_{\text{Salary}}(i)}{|\Omega|} = \frac{\sum_{r \in \Omega} \text{Salary}(r)}{|\Omega|} \tag{24}$$

As per (23) and (24), the average of a random variable X in probability theory and simple averages of an OLAP query provide the same outcome. Hence, we say that an average query in OLAP corresponds to expected values in probability theory.

4) SEMANTIC CORRESPONDENCE BETWEEN OLAP CONDITIONAL AVERAGES AND SR

The conditional average queries in OLAP calculate averages of a column with a WHERE clause. For example, we have an average SQL query with some conditions where the target column is numerical and conditional variables have arbitrary values. We have *OLAP Query:*

$$SELECT\ AVG\ (Salary)\ FROM\ Table\ 8$$
$$WHERE\ City = Seattle\ AND\ Profession = IT; \tag{25}$$

In probability theory, we compute the conditional average of a random number using its conditional expectation. For example, as per Def. 8, the conditional expectation of a random number Y with condition X is given as:

$$\mathsf{E}(Y|X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(Y = i_n|\ X)$$
$$f(i) = E(Y = i_n|X) \tag{26}$$

Here, the value $E(Y = i_n|X)$ is dependent on the value of i. Therefore, we say that $E(Y = i_n|X)$ is a function of i, which is given in (26). We compare the conditional expected value of $E(Y = i_n|X)$ with the output of the conditional AVG query in OLAP. We have *OLAP Query:*

$$SELECT\ AVG\ (Salary)\ FROM\ Table\ 8$$
$$WHERE\ City = Seattle\ AND\ Profession = IT;$$
$$\text{Conditional Expected Value: } \mathsf{E}(Salary|City = Seattle \cap Profession = IT) \tag{27}$$

$$\mathsf{E}(Y|X) = \sum_{i \in I_C} i \cdot \mathsf{P}(\mathrm{Y} = i\ |\ \mathrm{X}) \tag{28}$$

As per (27) and (28), the average of a random variable Y with condition X (Conditional Expected values) and the conditional average of an OLAP query provide the same outcome. Hence, we can say that a conditional average query in OLAP corresponds to the conditional expected values in probability theory. In Fig. 5, we demonstrate the semantic correspondence between the features of SR, OLAP, and ARM. At the top level, we consider OLAP and its features. In the middle, we have probability theory and its features, which work as the middle layer between OLAP, ARM and at the bottom layer, we provide ARM and its measures. In OLAP, we have conditional averages over binary columns,

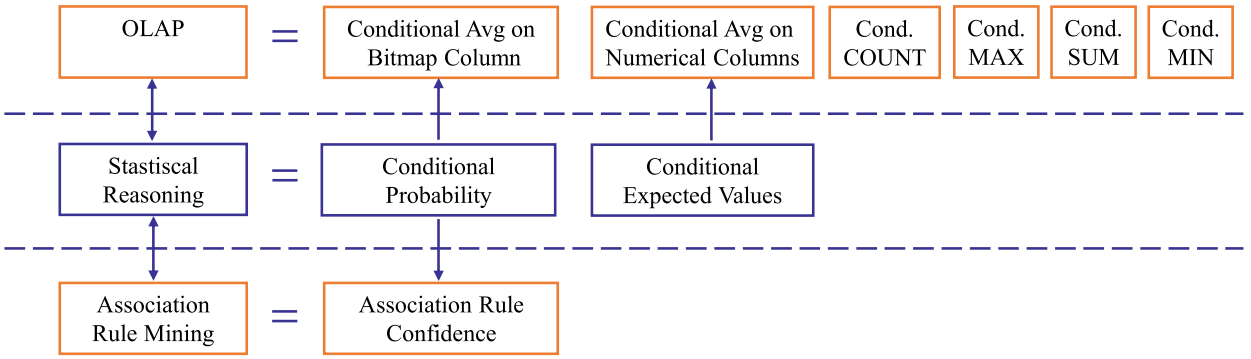


FIGURE 5. Demonstration of semantic correspondence between statistical reasoning, OLAP and association rule mining.

TABLE 9. Semantic correspondence between statistical reasoning, OLAP and association rule mining.

Concepts	Statistical Reasoning	OLAP	Association Rule Mining
Background	Probability Space	Database Table	Transaction Dataset
Data Notion	$(\Omega, \Sigma, \mathsf{P})$	$\{C_1 \times C_2 \ldots \times C_n\}$	$I = \{i_1, i_2 \ldots i_n\}$
Data Implementation	Table	Table	Bitmap Table *(in classical ARM)* and Table with discrete value columns *(in practical ARM tools)*
Average of a bitmap column $Y : \{0,1\}$ under dicing w.r.t setting bitmap columns $X_1 \ldots X_m$ to *truth values* (i.e., 1)	Conditional Probability $\mathsf{P}(Y\|X_1,\ldots,X_m) = \frac{\mathsf{P}(Y \cap X_1 \cap \ldots \cap X_m)}{\mathsf{P}(X_1 \cap \ldots \cap X_m)}$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = 1$ AND ... AND $X_m = 1$	Confidence $Conf(X_1, \ldots, X_m \Rightarrow Y)$ *(in classical ARM)*
Average of a bitmap column $Y : \{0,1\}$ under dicing w.r.t setting discrete columns $X_1 \ldots X_m$ to *arbitrary values*	Conditional Probability $\mathsf{P}(Y\|X_1{=}x_1,\ldots,X_m = x_m) = \frac{\mathsf{P}(Y \cap X_1{=}x_1 \cap \ldots \cap X_m{=}x_m)}{\mathsf{P}(X_1 \cap X_1{=}x_1 \cap \ldots \cap X_m{=}x_m)}$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = x_1$ AND ... AND $X_m = x_m$	Confidence $Conf(X_1{=}x_1, \ldots, X_m = x_m \Rightarrow Y)$ *(in practical ARM tools)*
Average of a numerical column $Y : \{i_0, \ldots, i_k\} \subseteq \mathbb{R}$ under dicing w.r.t setting discrete columns $X_1 \ldots X_m$ to *arbitrary values*	Conditional Expected Value $\mathsf{E}(Y\|X_1,\ldots,X_m) = \sum_{n=0}^{k} i_n \cdot \mathsf{P}(Y{=}i_n\|X_1{=}x_1,\ldots,X_m{=}x_m)$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = x_1$ AND ... AND $X_m = x_m$	-
Tools	SPSS, SAS, R	IBM Cognos, Palo, Mondrian, OLAP server, Pivot Table	Rapidminer, Orange, Weka

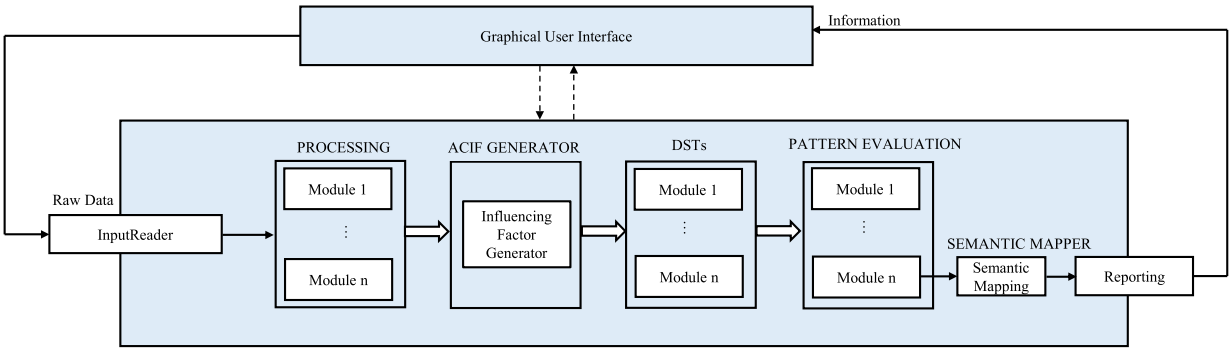


FIGURE 6. A high level abstraction of the framework for the unification of decision support techniques.

conditional averages over numerical columns, and different other conditional aggregates like Max, Min, Sum, etc. In OLAP, conditional averages on binary columns correspond to conditional probability, and they also correspond to confidence in ARM. However, conditional averages on numerical columns in OLAP correspond to conditional expected values

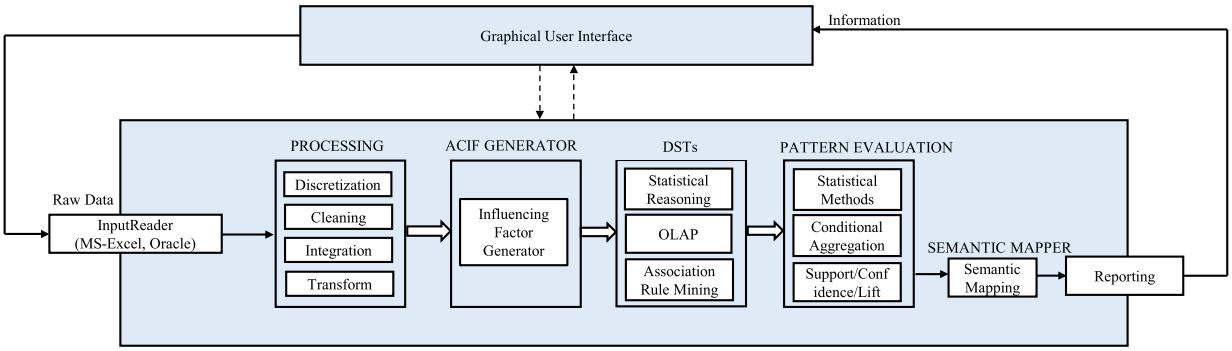


FIGURE 7. A detailed overview of the framework for the unification of decision support techniques.

in probability theory. Based on these semantic correspondences between SR, OLAP, and ARM, we are convinced that DSTs have common features with different names. However, they are being used differently. Therefore, the unification of SR, OLAP, and ARM will provide an advanced novel framework for next-generation decision support tools. In Table 9, we provide a list of semantic correspondence between the features of SR, OLAP, and ARM.

V. THE FRAMEWORK, EVALUATION AND EXPERIMENTS

In this section, the framework for the unification of three DSTs is presented. As a data science process provided by Schutt and O'Neil [66], the proposed framework is modular in design and every module in the framework can be displaced. In Fig. 6, we illustrate the high-level abstraction of the framework and based on the process of knowledge discovery in databases (KDD) [36], a detailed overview of the proposed framework is given in Fig. 7.

The framework consists mainly of seven major components. The Graphical User Interface (GUI) allows decision-makers to communicate with the framework to process the raw data. The data pre-processing includes various operations and checks, including discretization, cleaning, e.g., checking for corrupt data, reviewing the types of data, transforming and integrating data in useful formats, etc. The ACIF generator in the framework is developed for decision-makers to select the target columns and influencing factors to generate different combinations of data items. The decision support engine is a set of multiple DSTs, allowing decision-makers to select one or more techniques to process the data and get insights. The Pattern evaluation is used to find interesting information using different methods from SR, OLAP, and ARM. The semantic mapper is a manual process to map the results of DSTs and reports different semantic correspondences between them. A brief description of all the significant components of the proposed framework is given in Table 10.

A. IMPLEMENTATION OF THE PROPOSED FRAMEWORK

To demonstrate the usability of the proposed framework, an instance of the framework is developed using ASP.NET, an open-source framework for developing web applications. The resulting tool is an example of a next-generation decision support tool implemented by adopting the proposed framework. A summary of technologies and framework used for the implementation of the tool is given in Table 11. The programming code and other instructions on how to use the proposed tool are available in the GitHub repository [15]. The AJAX request methods are used throughout the tool's implementation to establish a connection between the client and server. JSON serialization and deserialization functions convert .NET objects (strings) to JSON format and JSON format to .NET objects. We use Oracle database and Microsoft Excel as databases and for OLAP, we have used relational algebra in the tool.

The tool first recognizes different kinds of data (discretized, numerical, categorical) and then develops generalized association rules for the various combinations of influencing factors and target columns. In the tool, if the selected target column is numerical, then the aggregate function is used, and the average value of the target column is calculated against the chosen influencing factors by the following SQL query; *Select AVG (target column) from table group by influencing factors*. If the specified target column is numerical, the aggregate function is employed in the tool, and the average value of the target column is determined against the chosen influencing factors using the SQL statement; *Select AVG (target column) from table group by influencing factors*. If the selected column is categorical, the tool uses the following SQL query to determine the conditional probability of the target column; *Select conditional probability of target column under influencing factor from table group by target column and influencing factors*. Both support and lift are calculated for numerical and categorical target columns. For the numerical target column, the order of columns is support, lift, an average value of the target column, and then influencing factors. For the categorical target column, the columns are listed in the following order: support, lift, conditional probability, target column, and influencing variables.

1) ACIF GENERATOR

In the tool, we have developed a function for ACIF generator and implemented it in the proposed framework.

TABLE 10. Summary of the components used to develop the framework for the unification of DSTs.

Component	Objective
Graphical User Interface	GUI is Used to communicate between decision-makers and the framework
Data pre-processing	The data pre-processing step includes various operations and checks, including discretization and cleaning (e.g., checking for corrupt data, reviewing the types of data, transforming and integrating data in useful formats, etc.)
ACIF	The ACIF generator is used to compute all the combinations of influencing factors
Decision Support Engine (DSE)	DSE is a set of DSTs used in the framework
Pattern Evaluation	In pattern evaluation, a set of measures from different DSTs are used to evaluate the usefulness of the patterns between the data items
Semantic Mapper	Semantic mapper is a process to compare the outcome of DSTs
Reporting	The reporting process is used to generate various reports for the outcome of DSTs

TABLE 11. Summary of the technologies and framework used for the implementation of the tool.

Description	Technologies
Programming Language	C#
Development Framework	ASP.NET
Requesting data through the web server	AJAX
Data access to the Oracle database	Oracle Data Provider (ODP), ODP.NET
Data access to the Microsoft Excel file	OLE DB

The ACIF generator is developed to select the target column and influencing factors to generate all possible combinations of the selected target column and influencing factors. First, the generator identifies the column combinations from the dataset and generates reports for the target column and influencing factors. The pseudo-code for the ACIF generator and ACIF report generator is given in Listing 1. In the pseudo-code, the *CREATE_COMBINATIONS* function is defined to pass the information of influencing columns and the number of columns. This function calculates the possible combinations of the selected influencing factors. In line 15, the *GENERATE_REPORT* function is defined to generate the reports for various combinations of influencing factors against target columns. This function passes the information about the table name, target columns and influencing columns. In this function, the SQL statement is used to retrieve the support, lift, conditional averages and influencing factors from the data source.

2) MATHEMATICAL DESCRIPTION OF THE ACIF GENERATOR

Let T be a database Table with multiple columns $C = \{X_1 : T_1, \ldots, X_n : T_n\}$, where $X_1 \ldots X_n$ represent column names and $T_1 \ldots T_n$ represent column types. To calculate various operations of SR, OLAP and ARM for T,

$$\forall 1 \leq \psi \leq n$$
$$\forall D = \{X'_1 : D_1, \ldots, X'_{\psi-1} : D_{\psi-1}\} \subseteq C(D_i = d_1, \ldots, d_{ni})$$
$$\forall d'_1 \in D_1, \ldots, d'_{\psi-1} \in D_{\psi-1}$$

Here, D is the subset of C and the influencing factor.

$$\forall Y : \mathbb{R} \in C \text{ or } Y = X_{ij} : B, X_i : d_i \in C$$

Y is the target column, $\mathbb{R}$ is the real-valued numbers then.

Support:

$$\mathsf{P}(Y, X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1}) \tag{29}$$

Average:

$$\mathsf{E}(Y \mid X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1}) \tag{30}$$

Lift:

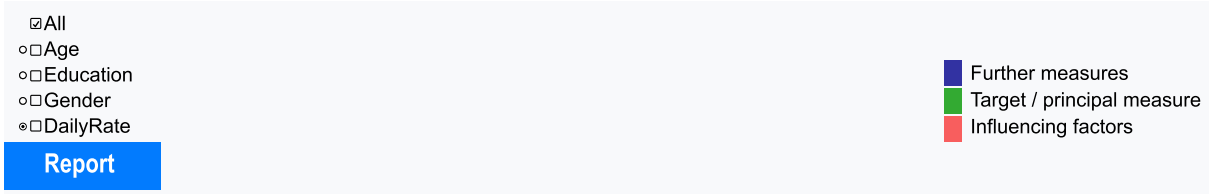
$$\frac{\mathsf{E}(Y \mid X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1})}{\mathsf{E}(Y)} \tag{31}$$

B. EXPERIMENTS ON THE PROPOSED FRAMEWORK

The experiment section demonstrates the potential of the introduced framework. The tool is evaluated on two real datasets and one synthetic dataset. The tool is tested on a computer with an Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz, 1800 Mhz, 4 Core(s), 8 Logical Processor(s), 16 GB RAM and Windows 10 x 64 operating system. The programming code, datasets, and other necessary instructions about the tool are available in the GitHub repository [15].

The datasets are summarized in Table 12, in which we highlight the number of records, number of attributes, and number of numeric attributes. The first Dataset, New Jersey (NJ) School Teacher Salaries (2016) [67] contains 138, 715 records and 15 attributes, while another real dataset, DC public government employees [68] contains 33, 424 records, which are huge in numbers to check the performance of the tool. In the table, we have described the dataset attributes with their types. Dataset NJ Teacher Salaries (2016) consists of salary, job, and experience data for the teachers and employees in New Jersey schools. The data are sourced from the (NJ) Department of Education. The second real dataset is a list of DC public government employees and their salaries in 2011. The second data set is sourced from the washington times via freedom of information act (FOIA) requests. We have also tested the proposed tool on the sample dataset UDS1 [69]. This dataset contains 1, 470 records with different combinations of numerical, categorical, and discretized attributes. A feature list obtained by parsing the UDS1 dataset is summarized in Table 13.

In the datasets, the target column is the one for which we are computing ARM operations, i.e., support, confidence, lift and OLAP averages with respect to an influencing factor. An influencing factor is an attribute that impacts



#	SUPPORT	LIFT	AVG_DailyRate	Age	Education	Gender
1	0.222	0.97	825.44	20-30		
2	0.116	0.98	822.42		A	
3	0.400	0.99	808.27			Female
4	0.056	0.96	837.75	20-30	A	
5	0.082	0.93	836.38	20-30		Female
6	0.018	0.86	928.96	20-30	A	Female
7	0.041	0.96	836.97		A	Female

FIGURE 8. A sample report comparing the results of OLAP and ARM measures is as follows: the ARM operations (support, confidence, lift) and OLAP operations (averages) are displayed. A sample dataset is used to generate the report, which includes all possible combinations of influencing factors and numerical target columns.

Listing 1 Pseudo-Code to Find the ACIF and Generate ACIF Reports

```
1:  function CREATE_COMBINATIONS(influencing_Columns[], numberofColumns)
2:      if numberofColumns == \text{0}~then
3:          return []
4:      return_Values = []
5:      for i = \text{1}~to LENGTH(influencing_Columns) do
6:          colName = influencingColumns[i]
7:          partialLst = REMOVE_COLUMN(i,influencingColumns)
8:          for each: j in CREATE_COMBINATIONS(partialLst, numberofColumns - 1) do
9:              APPEND_TO(return_Values,ADD_FIRST(colName,j))
10:         end for
11:     end for
12:     return return_Values
13: end function
14:
16: function GENERATE_REPORT(table_Name, target_Column, influencing_Columns[])
16:     for i = \text{1}~to LENGTH(influencing_Columns) do
17:         column_Combination = call:CREATE_COMBINATIONS(influencingColumns,i)
18:         for each: Combinations in column_Combination
19:             "SELECT COUNT(*)/ (SELECT COUNT(*) FROM "+table_Name+") AS SUPPORT,
AVG("+target_Column+") AS LIFT,
AVG("+target_Column+") AS AVG_targetColumn, "+ Combinations +"
FROM "+table_Name +"
GROUP BY "+ Combinations +"
ORDER BY "+ Combinations;"
20:         end for
21:     end for
22: end function
```

the target columns. Therefore, we also denote the WHERE clause as an influencing factor for the target column in OLAP computations. The UDS1 dataset consists four columns with different data types; age is discretized, gender is categorical, education is categorical and DailyRate is numerical. The column Age has the age groups as $20-30$, $30-40$, etc.,

TABLE 12. Summary of datasets used to evaluate performance of the proposed tool.

Dataset Type	Title	Records	Total Attributes	Numerical Attribute
Real	NJ Teacher Salaries (2016) [67]	138715	15	5
Real	Public government employees	33424	6	2
Synthetic	UDS1	1470	4	1

TABLE 13. A summary of different attributes obtained by parsing the datasets.

S.No.	Feature	Type
1	Age	DISCRETIZED
2	DailyRate	NUMERICAL
3	Education	CATEGORICAL
4	Gender	CATEGORICAL

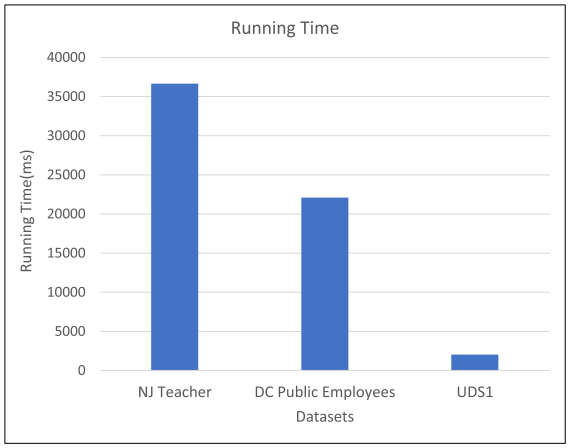


FIGURE 9. Running time and performance variation of the proposed tool induced by the number of records in the datasets.

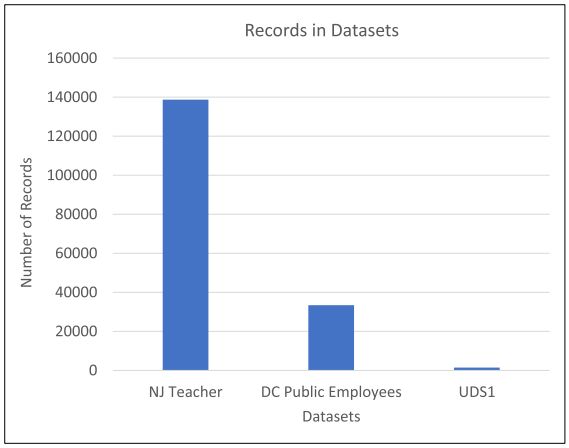


FIGURE 10. Number of records in datasets.

and gender has two categorical values; Male and female. Education has five categorical levels A, B, C, D, and E. For example, if we select education as the target column and its values are A, B, C, D, and E. Here, education is a factor and its values are instances of the factor. The tool calculates the conditional probability for each instance in the generated report. For example, suppose we select DailyRate as the target column and age, gender and education as influencing factors. In this case, all possible combinations of the target column are generated against all selected influencing factors.

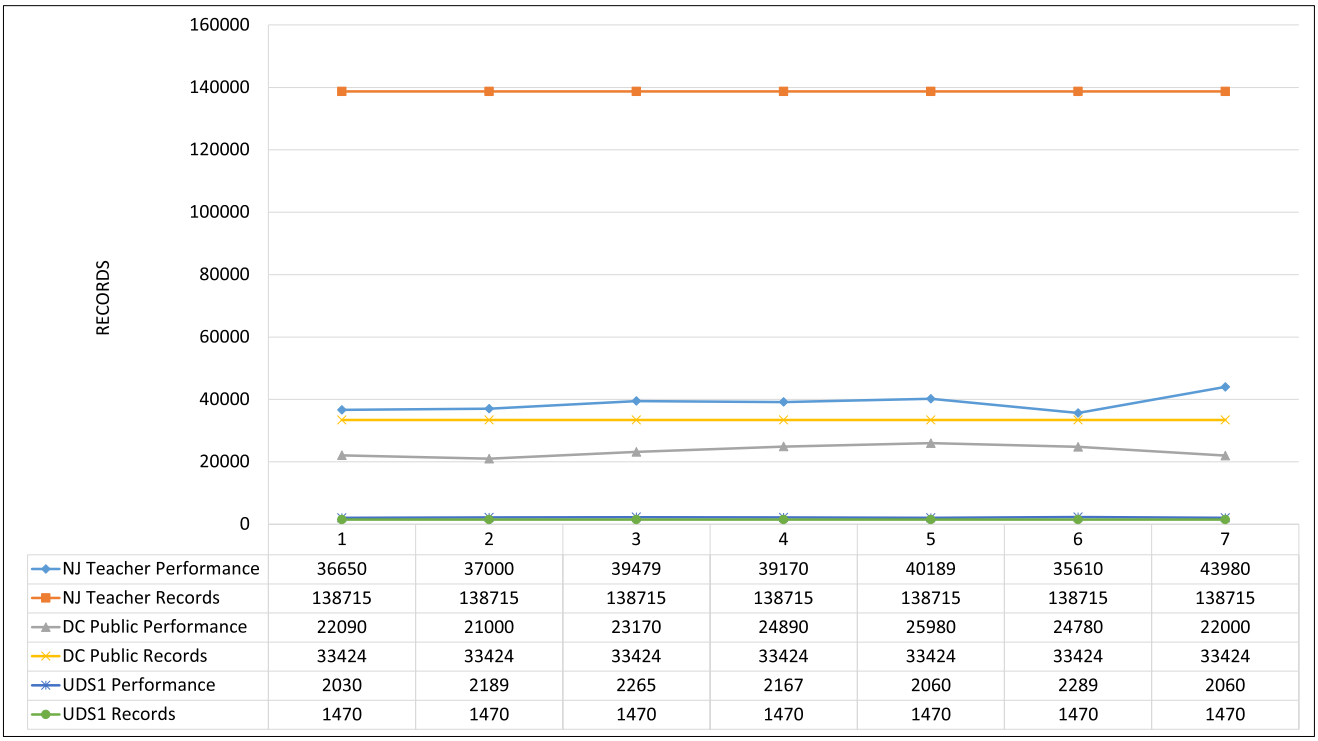
At the first step, the tool checks for the types of input data. Then it generates generalized association rules concerning the possible combinations of influencing factors and target columns. In the second step, the tool provides aggregate values, the conditional probability of the target column for each combination of influencing factors and target column. For SR, the tool calculates conditional probability and the mean value for the numeric target column concerning the influencing factors. For ARM operations, the tool calculates the support, confidence, and lift. For OLAP operations, the tool computes conditional averages. An overview of the computation of different SR, OLAP, and ARM operations is given in Fig. 8. In the report, the blue color code shows ARM operations. The green color code displays the target column, and the red color code indicates the influencing factors.

We have analyzed the performance of the proposed tool with three datasets. The performance of the tool varies with the number of records. If the number of records in a dataset is high, the tool has higher running time and slow performance. In Fig. 9, the performance variation induced by the number of records in a dataset is shown. The Dataset NJ Teacher has a huge number of records; therefore, its running time is $36,650$ milliseconds. Dataset DC Public Employees has $33,424$ records. Therefore, its running time is $22,090$ milliseconds, and the sample dataset UDS1 has a small number of records, i.e., $1,470$; therefore, its running time is $2,030$ milliseconds. Running time and performance variation of the proposed tool induced by the number of records in the datasets is shown in Fig. 10. A summary of records in datasets and performance variation of the tool with the datasets is given in Fig. 11.

C. ADVANTAGES OF THE PROPOSED TOOL OVER EXISTING DECISION SUPPORT TOOLS

In this section, we compare the capabilities of the proposed tool with one of the state of the art decision support platforms, i.e., RapidMiner [70].

Unlike any other decision support tool, the proposed tool altogether computes SR operations, i.e., conditional



	1	2	3	4	5	6	7
NJ Teacher Performance	36650	37000	39479	39170	40189	35610	43980
NJ Teacher Records	138715	138715	138715	138715	138715	138715	138715
DC Public Performance	22090	21000	23170	24890	25980	24780	22000
DC Public Records	33424	33424	33424	33424	33424	33424	33424
UDS1 Performance	2030	2189	2265	2167	2060	2289	2060
UDS1 Records	1470	1470	1470	1470	1470	1470	1470

FIGURE 11. Performance summary of the tool under two real datasets and one synthetic dataset.

#	SUPPORT	LIFT	CONDITIONAL_PROBABILITY	Education	Age
1	0.056	0.488	0.488	A	20-30
2	0.043	0.223	0.223	B	20-30
3	0.097	0.25	0.25	C	20-30
4	0.023	0.085	0.085	D	20-30
5	0.002	0.062	0.062	E	20-30

FIGURE 12. In the proposed tool: a sample project for generating all possible combinations of influencing factors against target columns.

probability, OLAP operation, i.e., conditional averages, and ARM operations, i.e., support, confidence, and lift, see Fig. 12. In addition, the tool computes the average value of a numerical target column against all possible combinations of influencing factors. In Fig. 8, a sample report is given for generating all possible combinations of influencing factors against the target column. However, in RapidMiner, to calculate the average value of a numerical target column against all possible combinations of influencing factors, a decision-maker needs to create multiple connections for all the possible combinations of influencing factors. Moreover, a decision-maker must create a new project for each dataset and repeatedly modify its columns and combinations. Therefore, in Fig. 13, we provide a sample use case to generate all possible combinations of influencing factors against the target column in RapidMiner.

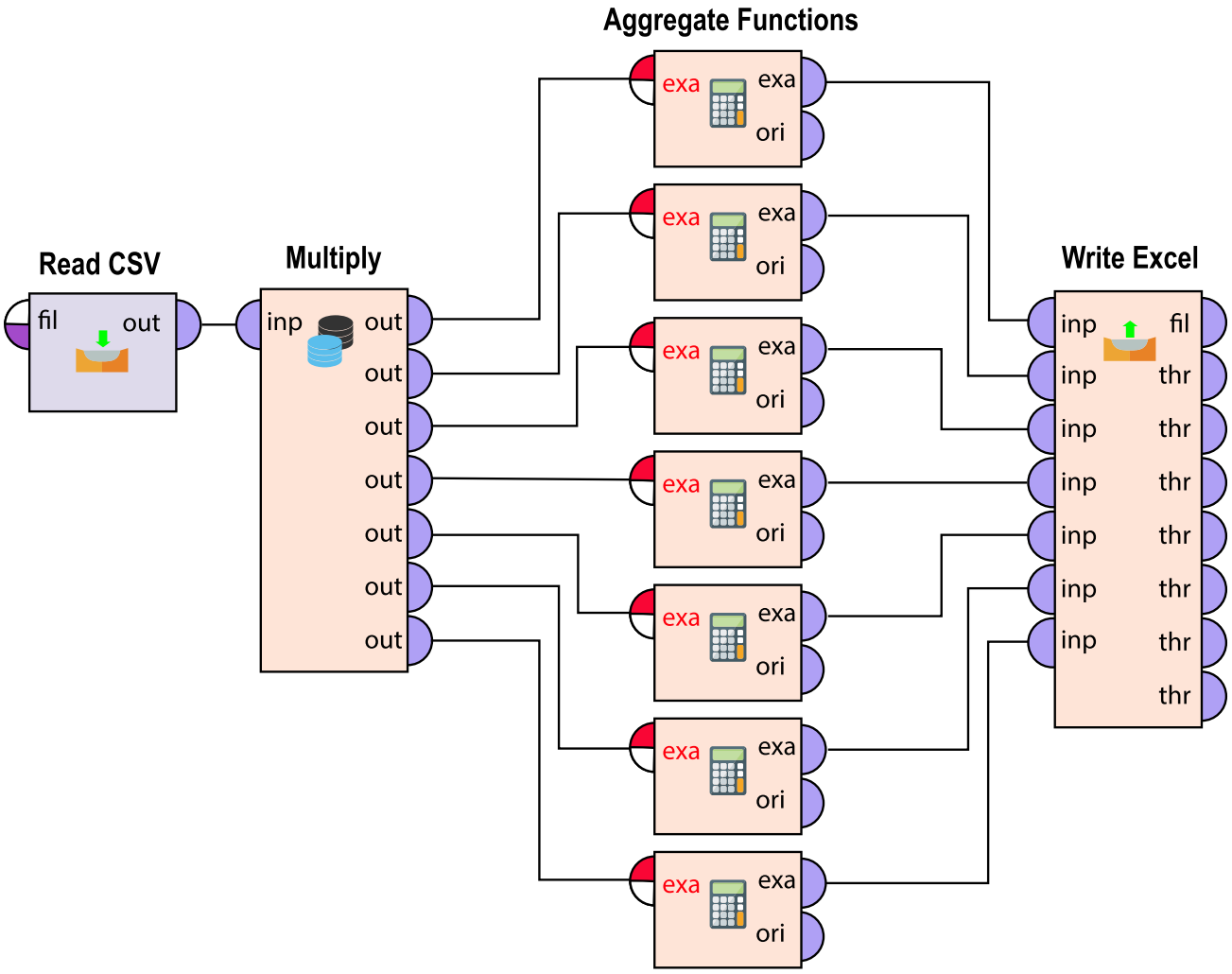


FIGURE 13. In RapidMiner: a sample project for generating all possible combinations of influencing factors against target columns.

TABLE 14. A sample list of premises and conclusions generated by RapidMiner for the influencing factors and target column.

Premises	Conclusion
Education = D Gender	Gender, Age = 30-40 Age = 30-40, Education = D

Additionally, in RapidMiner, the influencing factors and their values are stored in a single column called the 'conclusion' column as ''influencing factors=value''. The target column and its values are stored in the 'premises' column as ''Target Column=value''. A sample list of premises and conclusions generated by RapidMiner for the influencing factors and target column is displayed in Table 14. The representation of the target factors and influencing factors is difficult to understand. It is hard for decision-makers to identify each factor and its instance from the multiple tables. However, the proposed tool creates a separate column for each factor to identify the target column and influence factors quickly. In the tool, a decision-maker can select the target column and all influencing factors at once to generate all combinations of target factors and all influencing factors.

VI. FUTURE WORK

This paper provides a foundation for uncovering the semantic correspondences between DSTs and utilizing them to develop a framework for the unified usages of DSTs. However, the research is yet limited in scope to find the semantic correspondences between the three DSTs only; therefore, in the near future, more DSTs can be investigated to identify the semantic correspondences between them to develop cutting-edge frameworks for next-generation decision support tools. The unified usage of DSTs will not only be helpful in building robust frameworks for a variety of decision support tools but also open a new domain of research for hybrid DSTs.

Furthermore, we intend to build an advanced platform by implementing additional features in the proposed tool, e.g., Pearson correlation, regression, etc. We are also working on a new measure to identify any instance of Simpson's paradox in Big Data. Implementing these measures in the proposed platform will enable decision-makers to determine the genuine and unbiased impact factors.

The proposed tool has some performance issues with large datasets momentarily; therefore, we plan to scale up the performance of the tool by utilizing high-performance computing (HPC) infrastructure and making it available to the decision-makers. We intend to build it as a trustworthy platform and grow as a service provider in the near future.

VII. CONCLUSION

In this paper, we analyzed a series of approaches to overcome the divide between the three most popular DSTs, i.e., SR, OLAP and ARM. We contributed by elaborating the semantic correspondences between the foundations of SR, OLAP and ARM, i.e., probability theory, relational algebra and the itemset apparatus, respectively. The support of an itemset corresponds to the probability of a corresponding event and the confidence of an association rule corresponds to the conditional probability of two corresponding events. Furthermore, the OLAP average aggregate function corresponds to conditional expected values, which closes the loop between ARM, OLAP and probability theory with respect to the most important constructs in ARM and OLAP. We have proposed a novel framework for the unification of DSTs and implemented a tool to validate the concept of unification. The tool provides unified usage of DSTs in a classical decision support process and clarifies in how far the operations of SR, ARM, and OLAP can complement each other in understanding data, data visualization and decision making. The tool was developed on the basis of an open-source framework and tested with two real datasets and one synthetic dataset. The results and performance of the tool show valuable contributions towards developing the next-generation DSSs.

REFERENCES

[1] S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA, USA: Harvard Univ. Press, 1986.

[2] G. A. Gorry and S. M. S. Morton, "A framework for management information systems," Alfred P. Sloan School Manage., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 510-71, Feb. 1971.

[3] N. H. Nie, D. H. Bent, and C. H. Hull, *SPSS: Statistical Package for the Social Sciences*. New York, NY, USA: McGraw-Hill, 1970.

[4] *SAS User's Guide: Statistics; Version*, 5th ed., SAS Institute, Cary, NC, USA, 1987.

[5] E. Codd, S. Codd, and C. Salley, *Providing OLAP to User-Analysts: An IT Mandate*. East Falmouth, MA, USA: E. F. Codd and Associates, 1993.

[6] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[7] S. Chaudhuri and U. Dayal, "Data warehousing and OLAP for decision support," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1997, pp. 507–508, doi: 10.1145/253260.253373.

[8] Q. Wang, J. You, B. Zou, Y. Chen, X. Huang, and L. Jia, "Reduced quotient cube: Maximize query answering capacity in OLAP," *IEEE Access*, vol. 9, pp. 141524–141535, 2021.

[9] W. Thurachon and W. Kreesuradej, "Incremental association rule mining with a fast incremental updating frequent pattern growth algorithm," *IEEE Access*, vol. 9, pp. 55726–55741, 2021.

[10] H. Zhu, "On-line analytical mining of association rules," M.S. thesis, Brit. Columbia, Canada, School Comput. Sci., Simon Fraser Univ., British, CO, Canada, 1998.

[11] M. Kamber, J. Han, and J. Y. Chiang, "Metarule-guided mining of multi-dimensional association rules using data cubes," in *Proc. KDD 3rd Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1997, pp. 207–210.

[12] D. Draheim, "Future perspectives of association rule mining based on partial conditionalization (DEXA'2019 keynote)," in *Proc. DEXA 30th Int. Conf. Database Expert Syst. Appl.* in Lecture Notes in Computer Science, vol. 11706. Cham, Switzerland: Springer, 2019, pp. 40–49.

[13] G. Piatetsky. *Top Analytics, Data Science and Machine Learning Software*. Accessed: Dec. 21, 2021. [Online]. Available: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

[14] J. Han, Y. Fu, W. Wang, J. Chiang, O. R. Zaïane, and K. Koperski, "DBMiner: Interactive mining of multiple-level knowledge in relational databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1996, p. 550, doi: 10.1145/233269.280356.

[15] R. Sharma and S. A. Peious. *Towards Unification of Decision Support Technologies: Statistical Reasoning, OLAP and Association Rule Mining*. Accessed: Dec. 21, 2021. [Online]. Available: https://github.com/rahulgla/unification

[16] S. A. Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia, "Grand reports: A tool for generalizing association rule mining to numeric target values," in *Proc. DaWaK 22nd Int. Conf. Data Warehousing Knowl. Discovery* in Lecture Notes in Computer Science, vol. 12393. Cham, Switzerland: Springer, 2020, pp. 28–37.

[17] R. P. Salas, G. D. Edelson, P. S. Kleppner, and R. S. Shaver, "Data processing apparatus and method for a reformattable multidimensional spreadsheet," U.S. Patent 5 317 686, May 31, 1994.

[18] H. Wang, "Intelligent agent-assisted decision support systems: Integration of knowledge discovery, knowledge analysis, and group decision support," *Expert Syst. Appl.*, vol. 12, no. 3, pp. 323–335, Apr. 1997.

[19] W. Fan, H. Lu, S. E. Madnick, and D. Cheung, "DIRECT: A system for mining data value conversion rules from disparate data sources," *Decis. Support Syst.*, vol. 34, no. 1, pp. 19–39, Dec. 2002.

[20] N. Bolloju, M. Khalifa, and E. Turban, "Integrating knowledge management into enterprise environments for the next generation decision support," *Decis. Support Syst.*, vol. 33, no. 2, pp. 163–176, Jun. 2002.

[21] J. H. Heinrichs and J.-S. Lim, "Integrating web-based data mining tools with business models for knowledge management," *Decis. Support Syst.*, vol. 35, no. 1, pp. 103–112, Apr. 2003.

[22] V. Cho and E. W. T. Ngai, "Data mining for selection of insurance sales agents," *Expert Syst.*, vol. 20, no. 3, pp. 123–132, Jul. 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0394.00235

[23] N. Jukić and S. Nestorov, "Comprehensive data warehouse exploration with qualified association-rule mining," *Decis. Support Syst.*, vol. 42, no. 2, pp. 859–878, Nov. 2006.

[24] R. Rupnik, M. Kukar, and M. Krisper, "Integrating data mining and decision support through data mining based decision support system," *J. Comput. Inf. Syst.*, vol. 47, no. 3, pp. 89–104, 2007.

[25] S. T. March and A. R. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decis. Support Syst.*, vol. 43, no. 3, pp. 1031–1043, Apr. 2007.

[26] Z. Shi, Y. Huang, Q. He, L. Xu, S. Liu, L. Qin, Z. Jia, J. Li, H. Huang, and L. Zhao, "MSMiner—A developing platform for OLAP," *Decis. Support Syst.*, vol. 42, no. 4, pp. 2016–2028, Jan. 2007.

[27] N. Di Domenica, G. Mitra, P. Valente, and G. Birbilis, "Stochastic programming and scenario generation within a simulation framework: An information systems perspective," *Decis. Support Syst.*, vol. 42, no. 4, pp. 2197–2218, Jan. 2007.

[28] M. Charest, S. Delisle, O. Cervantes, and Y. Shen, "Bridging the gap between data mining and decision support: A case-based reasoning and ontology approach," *Intell. Data Anal.*, vol. 12, no. 2, pp. 211–236, Apr. 2008.

[29] Z. Y. Zhuang, L. Churilov, F. Burstein, and K. Sikaris, "Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners," *Eur. J. Oper. Res.*, vol. 195, no. 3, pp. 662–675, Jun. 2009.

[30] S. Liu, A. H. B. Duffy, R. I. Whitfield, and I. M. Boyle, "Integration of decision support systems to improve decision support performance," *Knowl. Inf. Syst.*, vol. 22, no. 3, pp. 261–286, Mar. 2010.
[31] Y. Peng, Y. Zhang, Y. Tang, and S. Li, "An incident information management framework based on data integration, data mining, and multi-criteria decision making," *Decis. Support Syst.*, vol. 51, no. 2, pp. 316–327, 2011.
[32] H. Ltifi, C. Kolski, M. B. Ayed, and A. M. Alimi, "A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases," *J. Decis. Syst.*, vol. 22, no. 2, pp. 69–96, Apr. 2013, doi: 10.1080/12460125.2012.759485.
[33] J. Dong, H. S. Du, S. Wang, K. Chen, and X. Deng, "A framework of web-based decision support systems for portfolio selection with OLAP and PVM," *Decis. Support Syst.*, vol. 37, no. 3, pp. 367–376, Jun. 2004, doi: 10.1016/S0167-9236(03)00034-4.
[34] I. Fister and I. Fister, Jr., "UARMSolver: A framework for association rule mining," 2020, *arXiv:2010.10884*.
[35] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C.-N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, May 2022, doi: 10.1145/3447772.
[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1999.
[37] H. X. Li and L. D. Xu, "Feature space theory – a mathematical foundation for data mining," *Knowl.-Based Syst.*, vol. 14, nos. 5–6, pp. 253–257, Aug. 2001, doi: 10.1016/S0950-7051(01)00103-4.
[38] Y. Zhu, C. Bornhövd, D. Sautner, and A. P. Buchmann, "Materializing web data for OLAP and DSS," in *Proc. WAIM 1st Int. Conf. Web-Age Inf. Manage.* Berlin, Germany: Springer, 2000, pp. 201–214.
[39] K. Gandhi, B. Schmidt, and A. H. C. Ng, "Towards data mining based decision support in manufacturing maintenance," *Proc. CIRP*, vol. 72, pp. 261–265, Jan. 2018.
[40] T. Imielinski, L. Khachiyan, and A. Abdulghani, "Cubegrades: Generalizing association rules," *Data Mining Knowl. Discovery*, vol. 6, no. 3, pp. 219–257, 2002.
[41] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules," *ACM SIGMOD Rec.*, vol. 27, no. 2, pp. 13–24, 1998.
[42] L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang, "Optimization of constrained frequent set queries with 2-variable constraints," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1999, pp. 157–168, doi: 10.1145/304182.304196.
[43] K.-N. T. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut, "Multidimensional association rules in Boolean tensors," in *Proc. SDM 11th SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2011, pp. 570–581.
[44] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, Mar. 1997, doi: 10.1145/248603.248616.
[45] Q. Chen, U. Dayal, and M. Hsu, "An OLAP-based scalable web access analysis engine," in *Proc. DaWak 2nd Int. Conf. Data Warehousing Knowl. Discovery*. Berlin, Germany: Springer-Verlag, 2000, pp. 210–223.
[46] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut, "Closed patterns meet n-ary relations," *ACM Trans. Knowl. Discovery From Data*, vol. 3, no. 1, pp. 1–36, Mar. 2009, doi: 10.1145/1497577.1497580.
[47] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decis. Support Syst.*, vol. 33, no. 2, pp. 111–126, 2002.
[48] J. W. Tukey, "Exploratory data analysis," in *Addison-Wesley Series in Behavioral Science*. Reading, MA, USA: Addison-Wesley, 1977.
[49] D. Donoho, "50 years of data science," *J. Comput. Graph. Statist.*, vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.
[50] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 1–12, Jun. 1996, doi: 10.1145/235968.233311.
[51] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. Ben Yahia, and D. Draheim, "On the potential of numerical association rule mining," in *Proc. FDSE 7th Int. Conf. Future Data Secur. Eng.* in Lecture Notes in Computer Science, vol. 12466. Singapore: Springer, 2020, pp. 3–20.
[52] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim, "A systematic assessment of numerical association rule mining methods," *Social Netw. Comput. Sci.*, vol. 2, no. 5, pp. 1–13, Sep. 2021.
[53] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, pp. 1–32, Sep. 2006, doi: 10.1145/1132960.1132963.
[54] C. D. Larose and D. T. Larose, *Discovering Knowledge in Data*. Hoboken, NJ, USA: Wiley, 2014, ch. Association Rules, pp. 247–265. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118874059.ch12
[55] R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim, "Expected vs. unexpected: Selecting right measures of interestingness," in *Proc. DaWaK 22nd Int. Conf. Data Warehousing Knowl. Discovery* in Lecture Notes in Computer Science, vol. 12393. Springer, 2020, pp. 38–47.
[56] B. Liu, W. Hsu, and S. Chen, "Using general impressions to analyze discovered classification rules," in *Proc. KDD 3rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. Palo Alto, CA, USA: AAAI Press, 1997, pp. 31–36.
[57] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining minimal non-redundant association rules using frequent closed itemsets," in *Proc. CL 1st Int. Conf. Comput. Log.* Berlin, Germany: Springer, 2000, pp. 972–986.
[58] R. J. Hilderman and H. J. Hamilton, *Knowledge Discovery and Measures of Interest. The Springer International Series in Engineering and Computer Science*. New York, NY, USA: Springer, 2001.
[59] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in *Proc. VLDB 21th Int. Conf. Very Large Data Bases*, 1995, pp. 420–431.
[60] H. Lu, L. Feng, and J. Han, "Beyond intratransaction association analysis: Mining multidimensional intertransaction association rules," *ACM Trans. Inf. Syst.*, vol. 18, no. 4, pp. 423–454, 2000, doi: 10.1145/358108.358114.
[61] I. Fister and I. Fister, "Association rules over time," 2020, *arXiv:2010.03834*.
[62] P. Fournier-Viger, J. Li, J. C.-W. Lin, T. T. Chi, and R. U. Kiran, "Mining cost-effective patterns in event logs," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105241. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119305581
[63] M. Shahin, S. A. Peious, R. Sharma, M. Kaushik, S. B. Yahia, S. A. Shah, and D. Draheim, "Big data analytics in association rule mining: A systematic literature review," in *Proc. 3rd Int. Conf. Big Data Eng. Technol. (BDET)*, Jan. 2021, pp. 40–49, doi: 10.1145/3474944.3474951.
[64] P. Y. Taser, K. U. Birant, and D. Birant, "Multitask-based association rule mining," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 933–955, 2020.
[65] K. E. Iverson, *A Programming Language*. Hoboken, NJ, USA: Wiley, 1962.
[66] R. Schutt and C. O'Neil, *Doing Data Science: Straight Talk From the Frontline*. Sebastopol, CA, USA: O'Reilly Media, 2013.
[67] S. Naik. (2016). *NJ Teacher Salaries*. [Online]. Available: https://data.world/sheilnaik/nj-teacher-salaries-2016
[68] M. Kalish. (2011). *DC Public Employee Salaries*. [Online]. Available: https://data.world/codefordc/dc-public-employee-salaries-2011
[69] R. Sharma and S. A. Peious. (2020). *UDS1*. [Online]. Available: https://github.com/rahulgla/unification/blob/master/UDS1.xlsx
[70] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 935–940.

RAHUL SHARMA (Graduate Student Member, IEEE) received the B.Tech. and M.Tech. degrees in computer science engineering from Dr. A. P. J. Abdul Kalam Technical University, India. He is currently pursuing the Ph.D. degree in computer science engineering with the Information Systems Group, Tallinn University of Technology, Estonia. He was an Assistant Professor with the Department of Information Technology, Ajay Kumar Garg Engineering College, Ghaziabad, India. His research interests include association rule mining, data science, big data, machine learning, and deep learning.

MINAKSHI KAUSHIK received the B.Tech. and M.Tech. degrees in computer science engineering from Dr. A. P. J. Abdul Kalam Technical University, India. She is currently pursuing the Ph.D. degree in computer science engineering with the Information Systems Group, Tallinn University of Technology, Estonia. Her research interests include association rule mining, big data, machine learning, and deep learning.

SIJO ARAKKAL PEIOUS received the master's degree in computer application from Annamalai University, India, in 2011, and the master's degree in e-governance technologies and services from the Tallinn University of Technology, Estonia, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Software Science. His research interests include association rule mining, big data, machine learning, and deep learning.

ALEXANDRE BAZIN received the Ph.D. degree from Université Pierre et Marie Curie, in 2014. He is currently working at the Lorraine Research Laboratory in Computer Science and its Applications (LORIA), Nancy, France, as a Postdoctoral Researcher. His research interests include lattice theory, symbolic approaches to pattern mining, and explainable artificial intelligence.

SYED ATTIQUE SHAH received the Ph.D. degree from the Institute of Informatics, Istanbul Technical University, Istanbul, Turkey. During his Ph.D. degree, he studied as a Visiting Scholar with the National Chiao Tung University, Taiwan, The University of Tokyo, Japan, and the Tallinn University of Technology, Estonia, where he completed the major content of his thesis. He worked as an Assistant Professor and the Chairperson at the Department of Computer Science, BUITEMS, Quetta, Pakistan. He was also engaged as a Lecturer at the Data Systems Group, Institute of Computer Science, University of Tartu, Estonia. His research interests include big data analytics, cloud computing, information management, and the Internet of Things.

IZTOK FISTER, JR. (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Maribor, Slovenia. He is currently an Assistant Professor at the University of Maribor. He has published more than 120 research articles in referred journals, conferences, and book chapters. His research interests include data mining, pervasive computing, optimization, and sport science. He has acted as a program committee member of more than 30 international conferences. Furthermore, he is a member of the editorial boards of five different international journals.

SADOK BEN YAHIA received the Habilitation degree to lead researches in computer sciences from the University of Montpellier, in 2009. His experience in teaching computer science and information systems is around 20 years. He was a Teaching Assistant with the Faculty of Sciences, Tunis, for two years, an Assistant Professor for seven years, and an Associate Professor for four years. Since 2013, he has been a Full Professor with the Faculty of Sciences. He has been a Professor with the Tallinn University of Technology (TalTech), since 2019. His research interests mainly include combinatorial aspects in big data and their applications to different fields, such as data mining, combinatorial analytics (maximum clique problem and minimal transversals), and smart cities (information aggregation and dissemination and traffic prediction).

DIRK DRAHEIM received the Ph.D. degree from Freie Universität Berlin and the Habilitation degree from Universität Mannheim, Germany. He is currently a Full Professor of information systems and the Head of the Information Systems Group, Tallinn University of Technology, Estonia. The Information Systems Group conducts research in large and ultra-large-scale IT systems. He is also an initiator and a leader of numerous digital transformation initiatives.