A probabilistic formalization of Association Rule Mining

Jose Antonio Ferez-Rubio, Carlos Perez-Vidal, Jose Vicente Segura-Heras

Departamento de Estadística, Matemáticas e Informática, Universidad Miguel Hernández de Elche

j.ferez@umh.es

Abstract. Data Mining has become an essential field in today's digital era, enabling organizations to extract actionable insights from ever-growing volumes of data. By uncovering patterns, trends, and hidden relationships within complex datasets, it supports informed decision-making and process optimization across domains ranging from business intelligence to biomedicine. Within this field, Association Rule Mining (ARM) stands out as a widely adopted technique for discovering meaningful relationships between variables in large datasets. This technique emerged in the early 1990s, when supermarkets began to have massive amounts of data on what products customers bought in each shopping basket. This huge amount of data was stored in what was called transaction databases. In 1993, Rakesh Agrawal, Tomasz Imieliński y Arun Swami published a paper where formally introduced the concept and problem of mining association rules in transactional data, laying the foundations of Association Rule Mining.[1] It should be noted that this paper was largely inspired by the previous contributions of Petr Hájek et al.[2] in 1960s, and Piatetsky-Shapiro [3] in 1991. Interestingly, none of these seminal papers formalizes a probabilistic approach to the theoretical foundations of ARM. However most of the books and scientific articles related to ARM employ inappropiate probabilistic approaches to define and evaluate core theoretical concepts and measures, such as support, confidence, or lift. These approaches are often applied inconsistently, lacking both methodological rigor and standardized definitions. This situation underscores the need to establish, with the aid of probability theory, a formal probabilistic theoretical framework based on the formalization of the foundations of ARM as presented in the seminal papers. This novel framework will allow, among other things, an improvement in the interpretability and comparability between the results of different studies, and will facilitates integration with other methodologies, such as statistical and machine learning techniques.

Keywords: Association Rule Mining; probability; formalization; ARM; Data Mining.

References

[1] R. Agrawal, T. Imieliński, A. Swami (1993). Mining Association Rules between Sets of Items in Large Databases, en P. Buneman et al., editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ACM Press, Washington, DC, USA, 207–216.

[2] P. Hájek, I. Havel, M. Chytil (1966). The GUHA method of automatic hypotheses determination. *Computing*, 1(4), 293–308.

[3] G. Piatetsky-Shapiro (1991). Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro, W. J. Frawley (Eds.), *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA, USA, 229–248.

Acknowledgment. Project CPP2021-008593, grant MICIU/AEI/10.13039/501100011033 of the Spanish Government and the European Union-NextGenerationEU/PRTR.

Indicar la preferencia (subrayar la opción elegida): [✓ póster] o charla.

Indicar la preferencia (subrayar la opción elegida): [✓ Lunes]/Martes o Jueves/Viernes.