A SEMANTIC APPROACH FOR POST-PROCESSING ASSOCIATION RULES IN THE CONTEXT OF BREAST CANCER IN BRAZIL

Paôla de Oliveira Souza
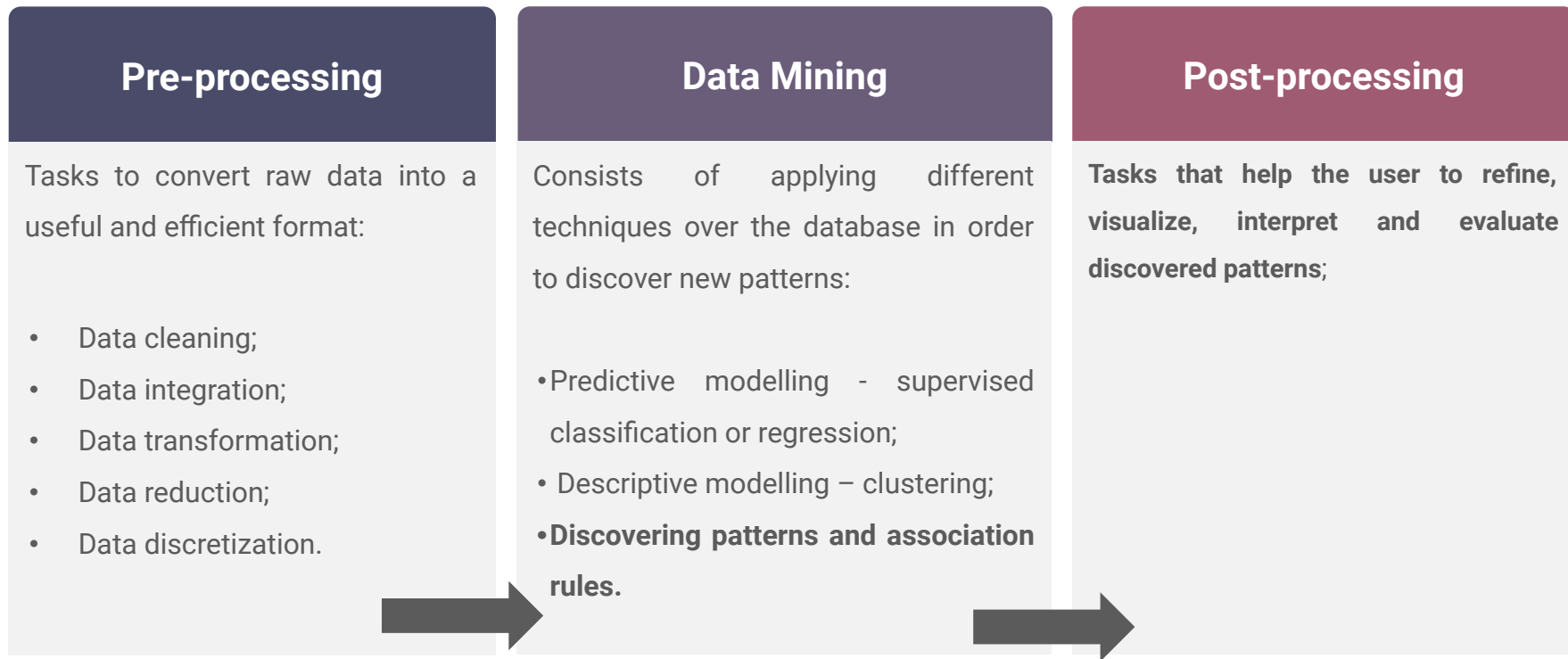
Aeronautics Institute of Technology

Advisor Professor Dr. José Maria Parente de Oliveira/ Professor Dr. Ricardo Santos

RESEARCH CONTEXT

KDD and Data Mining

Rogalewicz e Sika (2016) presented a general three-stage KDD model as follows:

Pre-processing	Data Mining	Post-processing
Tasks to convert raw data into a useful and efficient format: • Data cleaning; • Data integration; • Data transformation; • Data reduction; • Data discretization.	Consists of applying different techniques over the database in order to discover new patterns: • Predictive modelling - supervised classification or regression; • Descriptive modelling – clustering; • **Discovering patterns and association rules.**	**Tasks that help the user to refine, visualize, interpret and evaluate discovered patterns;**

Association Rules Mining (ARM)

ARM is one of the most famous DM techniques. It aims to discover patterns in transactional databases in the form of Association Rules (ARs).

Definition	ARs are implications of the form **IF Antecedent THEN Consequent** $$X \longrightarrow Y$$
Health Domain Application	The ARM can be widely used in the health domain because this domain produces a large volume of transactional data, most of them are patient history and treatment records (BYTYçI; AHMEDI; LISI, 2017).

Association Rules Mining (ARM)

However, ARM has two major challenges:

Challenge	• **Quantity:** a huge number of rules can be extracted; • **Quality:** not all rules are interesting.

Thes facts make the interpretation of the results very complicated and sometimes impossible for the domain expert.

Association Rules Mining (ARM)

Challenge

- One possible solution to reduce the complexity of user tasks in post-processing stage would be the creat methods for automatic evaluation of extracted rules.

- However, the automation of the DM process is difficult, since the the rules interestingness strongly depends on domain knowledge and user goals (NANDHINI; SIVANANDAM, 2020).

Semantic Data Mining

- **Semantic Data Mining** is a research field that deals with data mining tasks that systematically incorporate domain knowledge, represented in formal semantic models, so they can be processed by computers (SIRICHANYA; KESORN, 2021).

- **Ontologies** are the most advanced and complete formalism for representation of specialized knowledge to improve the process of knowledge discovery (SILVA; RIBEIRO, 2019).

- Despite being promising, the data mining process supported by ontologies is still in its initial stage. (IDOUDI et al.,2016).

Research Problem

Association Rules Mining

- ARM generates a large number of rules in which most are redundant, trivial, and conflicting with the domain knowledge;

- Domain experts have to browse and prune the results, looking for useful and unexpected patterns which becomes an impossible task when mining large databases.

- Thousands of rules are extracted from a database of several dozens of attributes and several hundreds of transactions.

Mining Health Data

- In Brazil a large volume of information can be found in the systems of the IT department of the Ministry of Health. However, there is still a lack of tools to transform that data into useful knowledge;

- Managers and health professionals miss the opportunity to provide better health care and improve the control of diseases with high prevalence and high mortality, such as breast cancer.

- One critical factor in mining health data is the use of generic DM techniques rather than using specialized solutions for this sort of data (RODRIGUES et al., 2022);

Research Problem

The development of DM systems, especially in complex domains such as the health area, requires efforts to create specialized methods and tools that incorporate domain knowledge, represented by an ontology, to automate the analysis of results, in order to reduce the complexity of users' tasks in post-processing and help them to discover interesting and useful patterns.

Research Objective

✓ The aim of this work is to create a semantic approach for post-processing association rules in the context of breast cancer in Brazil.

✓ This approach involves the development of semantic pruning tasks to eliminate redundant, trivial and inconsistent rules according to the application domain, represented by an ontology, and ranking the rules using a measure of semantic interestingness.

Conclusions

1. We confirmed from the literature review the importance of the development of semantic approaches to data mining. We also verified that semantic approaches for post-processing association rules in specific contexts have shown to be promising in advancing the production of useful knowledge from ARM.

1. The present research proposal can contribute to the advancement of research in semantic computational techniques for automating the analysis of association rules using ontologies.

1. In the context of breast cancer in Brazil, this research can also contribute to the production of useful knowledge from SUS data.

References

BYTYçI, E.; AHMEDI, L.; LISI, F. A. Enrichment of association rules through exploitation of ontology properties – healthcare case study. Procedia Computer Science, v. 113, p. 360–367, 2017. ISSN 1877-0509.

DOMINGUES, M.; REZENDE, S. Using taxonomies to facilitate the analysis of the association rules. In: . [S.l.: s.n.], 2005. 22,

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies, v. 43, n. 5, p. 907–928, 1995. ISSN 1071-5819.

GUARINO, N. The ontological level. In: . Philosophy and the Cognitive Science. Vienna: Holder-Pivhler-Tempsky, 1995. p. 443–456.

HAMANI, M.; MAAMRI, R. Ontology-driven method for ranking unexpected rules. In: . [S.l.: s.n.], 2009. v. 547. 24, 27

IDOUDI, R. et al. Ontology knowledge mining based association rules ranking. Procedia Computer Science, v. 96, p. 345–354, 2016. ISSN 1877-0509. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 20th International Conference KES-2016.

 MANSINGH, G.; OSEI-BRYSON, K.-M.; REICHGELT, H. Using ontologies to facilitate post-processing of association rules by domain experts. Information Sciences, v. 181, n. 3, p. 419–434, 2011. ISSN 0020-0255. Disponível em: https://www.sciencedirect-.com/science/article/pii/S0020025510004810>. 23, 25

 MARINICA, C.; GUILLET, F. Knowledge-based interactive postmining of association rules using ontologies. IEEE Transactions on knowledge and data engineering, IEEE, v. 22, n. 6, p. 784–797, 2010. 9, 13, 22, 23

References

NANDHINI, M.; SIVANANDAM, S. N. Extraction of significant association rules using pre- and post-mining techniques—an analysis. In: . Ontology-Based Information Retrieval for Healthcare Systems. John Wiley Sons, Ltd, 2020

OYELADE, O.N., EZUGWU, A.E. & ADEWUYI, S.A. Enhancing reasoning through reduction of vagueness using fuzzy OWL-2 for representation of breast cancer ontologies. *Neural Comput & Applic* **34,** 3053–3078 (2022).

PAUL, R. et al. Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain. Journal of biomedical semantics, v. 5, p. 8, 02 2014.

RODRIGUES, D. et al. Associative patterns in health data: exploring new techniques. Health and Technology, Springer, p. 1–17, 2022.

ROGALEWICZ, M.; SIKA, R. Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. Management and Production Engineering Review, Production Engineering Committee of the Polish Academy of Sciences, Polish Association for Production Management, n. No 4, 2016.

SIRICHANYA, C.; KESORN, K. Semantic data mining in the information age: A systematic review. International Journal of Intelligent Systems, v. 36, p. 3880 – 3916, 2021.