Mining Sematic Association Rules from RDF Data

Nima Khodadadi[*1], M. G. El-Mahgoub[2], Rokaia M. Zaki[3]

[1] Department of Civil and Architectural Engineering, University of Miami, Coral Gables, FL, USA
[2] Basic science department, Delta higher institute for engineering and technology, Mansoura, 35111, Egypt
[3] Higher Institute of Engineering and Technology, Kafrelsheikh, Egypt;
[3]Department of Electrical Engineering, Shoubra Faculty of Engineering, Benha University, Egypt
Emails: nima.khodadadi@miami.edu; melmahgoub@hotmail.com; rukaia.emam@feng.bu.edu.eg

Abstract

Many fields rely heavily on the accurate and consistent portrayal of structured data. In order to effectively express and link information on the Semantic Web, RDF (Resource Description Framework) data is essential. Here, we present a process for extracting semantic association rules from RDF data. For our method, we employ the Apriori algorithm to mine the RDF triples for hidden connections between ideas and relationships. Using metrics such as confidence, support, and lift, we examine how well our model performs. We also give visual representations, like as scatter plots and clustered matrices, to make the correlations easier to understand and analyse. The findings validate our model's potential to unearth significant relationships, which in turn reveal important details about the RDF data's underlying semantics. Our findings are discussed, and suggestions for further study are provided.

Keywords: RDF data; semantic association rules; mining; Apriori algorithm; confidence; support lift; visualizations; Semantic Web.

1. Introduction

There is an urgent demand for efficient methods to represent and analyze structured data due to the proliferation of digital information on the web. With its ability to express and link data on the web, the Resource Description Framework (RDF) has become an important standard for data integration and communication [1]. In order to describe knowledge, RDF uses subject-predicate-object triples to record the connections between various resources. Due to its adaptability across domains and its capacity to allow interoperability, RDF data is of critical value in conveying structured information. It has found widespread use in fields as diverse as e-commerce, healthcare, cultural preservation, and basic scientific inquiry. To better represent knowledge and facilitate sophisticated reasoning and data integration, businesses can take advantage of RDF's graph-based structure to semantically define their resources and establish relationships between them [2].

However, the vast amount of interconnected RDF data presents a challenge in uncovering valuable insights and extracting meaningful patterns. Mining semantic association rules from RDF data has emerged as a compelling approach to address this challenge. The motivation behind this research is twofold. Firstly, by mining semantic association rules from RDF data, we can reveal hidden associations and correlations among resources and their attributes. These associations provide valuable knowledge that can assist in decision-making, recommendation systems, ontology engineering, and knowledge discovery. Moreover, they can enhance data exploration, anomaly

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

detection, and query optimization [3-4]. Secondly, mining semantic association rules from RDF data enables the discovery of implicit knowledge and the identification of relationships that may not be readily apparent. RDF data often contains incomplete or sparsely represented information, making it challenging to fully exploit its potential. By leveraging association rule mining techniques, we can extract meaningful relationships and enrich the existing RDF data with additional semantic information. This process enhances the data's comprehensiveness, improves its quality, and enables more accurate reasoning and inference [5-6].

The potential applications of mining semantic association rules from RDF data are vast and span various domains. In e-commerce, for example, the extracted association rules can facilitate personalized product recommendations, cross-selling, and market basket analysis. In healthcare, they can aid in identifying co-occurring medical conditions or predicting adverse drug reactions [7-8]. Furthermore, in cultural heritage and scientific research, mining semantic association rules can help identify hidden connections between resources, enabling new discoveries and insights [9].

In this paper, we make several contributions to the field of mining semantic association rules from RDF data. Firstly, we propose a novel algorithm that combines association rule mining techniques with semantic reasoning to extract meaningful associations from RDF data. Our approach considers the graph-based nature of RDF and exploits the inherent semantics encoded in the data. Secondly, we investigate the enrichment of RDF data by integrating external knowledge sources.

The remainder of this paper is organized as follows. In Section 2, we provide a detailed background on RDF data representation and association rule mining techniques. Section 3 presents our proposed methodology for mining semantic association rules from RDF data, including the integration of semantic reasoning and external knowledge. In Section 4, we discuss the experimental setup and evaluation of our approach using real-world RDF datasets. The results and analysis are presented in Section 5. Finally, we summarize our findings, discuss the implications, and outline future research directions in Section 6.

2. Related Work

In this section, we review existing studies and research efforts related to mining association rules from RDF data. These investigations have aimed to leverage the underlying semantic relationships and knowledge encoded in RDF to extract valuable patterns and associations. For example, Kahani [1] proposed a new approach for mining semantic association rules in the context of stock market prediction. They applied association rule mining techniques to RDF data, considering the semantic relationships encoded in the data. Their work demonstrated the potential of leveraging semantic associations for financial prediction tasks. In addition, Barati et al [2] introduced a method for mining semantic association rules from RDF data, focusing on enhancing the discovery of interesting and meaningful patterns. They presented an algorithm that incorporated background knowledge and domain-specific ontologies to enrich the associations extracted from RDF data. The study demonstrated the effectiveness of their approach in discovering meaningful associations. In a related work, Barati et al [3] proposed the SWARM framework, which aimed to mine semantic association rules from semantic web data. The framework utilized a graph-based representation of RDF data and incorporated reasoning techniques to infer additional associations. Their research highlighted the potential of integrating reasoning mechanisms into association rule mining from RDF data. Furthermore, Alfrjani et al [5] proposed a novel approach to ontology-based semantic modeling for opinion mining. Although not directly related to association rule mining in RDF data, their work demonstrates the importance of leveraging semantic knowledge for extracting meaningful insights from textual data. In the field of ontology modeling, LePendu et al. [6] introduced the Ontology Database method, which aimed to semantically model brainwave data. Although their study does not focus on association rule mining, it highlights the significance of semantic modeling techniques and their potential applications in various domains. Nebot et al [7] focused on finding association rules in semantic web data. Their research explored techniques for discovering meaningful associations by leveraging the semantic relationships encoded in RDF data. The study demonstrated the potential of association rule mining in semantic web environments. While the studies [8-11] do not directly address mining semantic association rules from RDF data, they are relevant to the broader field of data mining, intelligent systems, and artificial intelligence. These studies emphasize the importance of leveraging advanced computational techniques for various applications and highlight the potential of extracting meaningful insights from different types of data.

Doi: https://doi.org/10.54216/JAIM.040105

Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

In this paper, we build upon the foundations laid by these studies and contribute to the field of mining semantic association rules from RDF data. We propose a novel algorithm that combines association rule mining techniques with semantic reasoning to extract meaningful associations. Additionally, we investigate the enrichment of RDF data by integrating external knowledge sources. Our approach aims to improve the quality and relevance of the discovered association rules and demonstrate their potential applications in different domains.

3. Methodology of our Solution

In this section, we present our proposed methodology for mining semantic association rules from RDF data. Our approach aims to leverage the inherent semantic relationships encoded in RDF to extract meaningful associations and patterns. We combine traditional association rule mining techniques with semantic reasoning and external knowledge integration to enhance the quality and relevance of the discovered rules.

A. Data Preparation

The first step in our methodology involves preprocessing and transforming the RDF data to prepare it for association rule mining. We consider the graph-based nature of RDF and the semantic relationships between resources. We convert the RDF data into a suitable format, such as a transactional representation, to facilitate subsequent mining operations. Additionally, we handle data cleaning tasks, including removing noise, handling missing values, and resolving redundancies, to ensure the integrity of the data [10-11].

B. Association Rule Mining

Once the RDF data is preprocessed, we apply association rule mining techniques to extract semantic associations. In our methodology, we employ the Apriori algorithm, a classical approach for mining association rules, to extract semantic associations from RDF data. The Apriori algorithm efficiently discovers frequent item sets and generates association rules based on the concept of itemset support. The Apriori algorithm follows a step-by-step procedure, given in Algorithm 1 showing in figure 1.

Doi: https://doi.org/10.54216/JAIM.040105

Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

Algorithm: Apriori

Input: Data D composed of transactions T. Lowest support threshold min_{sup} .

Output: Collection of repeated item sets and relating association rules.

Procedure:

Initialize:

1. Generate the set of frequent 1-itemsets by scanning the D and counting the occurrences of each item.
2. Prune occasional 1-itemsets by eliminating those below the min_{sup} threshold.
3. Repeat until no more repeated item sets can be generated:
4. Generate candidate item sets of length $k + 1$ from frequent item sets of length k.
5. Join the frequent item sets of length k to form candidate item sets by creating $(k + 1)$-itemsets.
6. Prune candidate item sets that contain subsets that are infrequent.
7. Scan the D to count the occurrences of each candidate itemset.
8. Support Counting Step: For each $t \in T$ in the database, check if the candidate itemset is a subset of the t.
9. Prune infrequent candidate item sets below the min_{sup} threshold.
10. Set the frequent item sets of length $k + 1$ as the remaining candidate itemsets.
11. Generate association rules from the frequent item sets:
12. For each frequent itemset, generate all possible non-empty proper subsets as potential antecedents.
13. For each potential antecedent, generate the corresponding consequent by subtracting the potential antecedent from the frequent itemset.
14. Calculate $confidence(M_1 \rightarrow M_2)$
15. Prune association rules that do not meet the minimum confidence threshold.
16. Return the set of frequent item sets and related association rules.

Figure 1: Apriori Algorithm

C. Semantic Reasoning

To enhance the mining process, we incorporate semantic reasoning techniques into our methodology. We leverage the ontological knowledge encoded in the RDF data and utilize reasoning mechanisms to infer additional relationships and expand the search space for association rule mining. By considering the semantic implications of the relationships between resources, we can uncover implicit associations that may not be directly represented in the data [13].

We employ reasoning engines, such as OWL reasoners or rule-based inference engines, to perform logical deductions and infer new knowledge based on the RDF data and the underlying ontologies. This step allows us to discover hidden associations and enrich the association rule mining process.

D. External Knowledge Integration

In addition to semantic reasoning, we integrate external knowledge sources to enhance the quality and relevance of the discovered association rules. We leverage domain-specific ontologies, background knowledge bases, or external data sources to enrich the RDF data and provide additional context for rule mining. By incorporating this external knowledge, we can capture domain-specific semantics and improve the comprehensiveness of the discovered associations.

We establish mappings between the RDF data and the external knowledge sources, aligning relevant concepts, properties, or instances. This integration allows us to leverage the complementary information and semantic relationships present in the external knowledge sources, leading to more accurate and meaningful association rules.

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

E. Evaluation Metrics

To evaluate the effectiveness and quality of the mined semantic association rules, we define appropriate evaluation metrics. We consider metrics such as support, confidence, lift, and interestingness measures to assess the significance and usefulness of the discovered rules. Support indicates how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. We calculate it as:

$$support(M) = \frac{\text{\# user watchlists containing M}}{\text{total number of user watchlists}} \quad (1)$$

where for Market Basket Optimization, it is computed as:

$$support(I) = \frac{\text{\# transactions containing I}}{\text{total number of transactions}} \quad (2)$$

Confidence, on the other hand, indicate how likely item B is bought when item A is bought, stated as $\{A -> B\}$. It is calculated as follows:

$$confidence(M_1 \rightarrow M_2) = \frac{\text{\# user watchlists containing } M_1 \text{ and } M_2}{\text{number of user watchlists containing } M_1} \quad (3)$$

whereas for Market Basket Optimization, it is computed as:

$$confidence(I_1 \rightarrow I_2) = \frac{\text{\# transactions containing } I_1 \text{ and } I_2}{\text{number of transactions containing } I_1} \quad (4)$$

Lift signifies how likely the item B is purchased when the item A is purchased while controlling for how popular item B is. it is computed as:

$$lift(M_1 \rightarrow M_2) = \frac{Confidence(M_1 \rightarrow M_2)}{Support(M_2)} \quad (5)$$

The evaluation phase involves measuring the efficiency, scalability, and accuracy of our approach on real-world RDF datasets. We analyze the performance of the mining process, the quality of the discovered rules, and the impact of incorporating semantic reasoning and external knowledge integration.

4. Evaluation and Case Studies

In our study, we conducted an experiment on a case study of Market Basket Analysis. Market Basket Analysis is a fundamental technique employed by large retailers to uncover associations between items based on customer transactions. Its purpose is to identify combinations of items that frequently occur together in transactions, enabling retailers to reveal relationships and patterns among the items that customers purchase. For our experiment, we utilized a dataset comprising 38,765 rows of purchase orders from various grocery stores. This dataset allowed us to analyze the transactions and generate association rules using Market Basket Analysis techniques [14]. The application of Market Basket Analysis to the dataset enabled us to derive association rules, which provide insights into the purchasing behaviors of customers. These rules allowed us to identify frequent item combinations and understand the relationships between different items in customers' shopping baskets. The analysis of association rules offers valuable information for retailers, helping them make informed decisions about product placement, cross-selling, and customer segmentation strategies. By conducting this case study on Market Basket Analysis, we aimed to demonstrate the effectiveness of our methodology for mining semantic association rules from RDF data. The insights gained from the experiment contribute to the understanding of customer purchasing behaviors and provide practical implications for improving retail operations and customer satisfaction.

In this section, we present the results of our proposed model for mining semantic association rules from RDF data. The performance and effectiveness of the model are evaluated in terms of various metrics, including confidence, support, and lift. Table 1 summarizes the key findings and highlights the most significant associations discovered.

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

Table 1: Summary of Association Rules and Metrics

	Item #1	Item #2	Support	Confidence	Lift
0	Instant food products	baby cosmetics	0.002747	0.035088	8.51462
1	Instant food products	bags	0.001374	0.017544	3.192982
2	Instant food products	liqueur	0.004121	0.052632	4.25731
3	abrasive cleaner	cleaner	0.005495	0.181818	4.564263
4	artif. sweetener	baby cosmetics	0.001374	0.035714	8.666667
5	artif. sweetener	cocoa drinks	0.002747	0.071429	3.25
6	artif. sweetener	cookware	0.002747	0.071429	3.058824
7	artif. sweetener	frozen chicken	0.001374	0.035714	5.2
8	artif. sweetener	honey	0.002747	0.071429	4
9	artif. sweetener	rubbing alcohol	0.001374	0.035714	5.2
10	artif. sweetener	specialty vegetables	0.002747	0.071429	4.727273
11	artif. sweetener	tea	0.004121	0.107143	3
12	baby cosmetics	canned fruit	0.001374	0.333333	11.55556
13	baby cosmetics	liquor	0.002747	0.666667	4.902357
14	baby cosmetics	liquor (appetizer)	0.001374	0.333333	3.851852
15	baby cosmetics	nut snack	0.001374	0.333333	11.55556
16	baby cosmetics	salt	0.002747	0.666667	6.14346
17	baby cosmetics	soups	0.001374	0.333333	5.275362
18	baby cosmetics	turkey	0.001374	0.333333	3.324201
19	bags	dental care	0.002747	0.5	11.375

To gain insights into the relationships and patterns captured by the association rules, we employ scatter plots. The scatter plot visualizations depict the confidence and support values of each rule, allowing us to identify interesting associations based on their strength and prevalence. Figure 2 a) illustrates a scatter plot of confidence against support for the association rules mined by our model. Each point on the scatter plot represents an association rule, with its position determined by the corresponding confidence and support values.

By analyzing the scatter plot, we can identify rules with high confidence and support, indicating strong and frequent associations. Additionally, we can observe patterns and trends, such as clusters or outliers, which may provide further insights into the relationships between different items or entities.

Figure 2 b) displayed a grouped matrix of association rules mined by our model. The rows and columns of the matrix represent the antecedents and consequents of the rules, respectively. Each cell in the matrix corresponds to a specific rule and provides information about its confidence, support, or other relevant metrics. By observing the patterns and clusters in the matrix, we can identify frequent associations and potential dependencies between different items or entities.

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

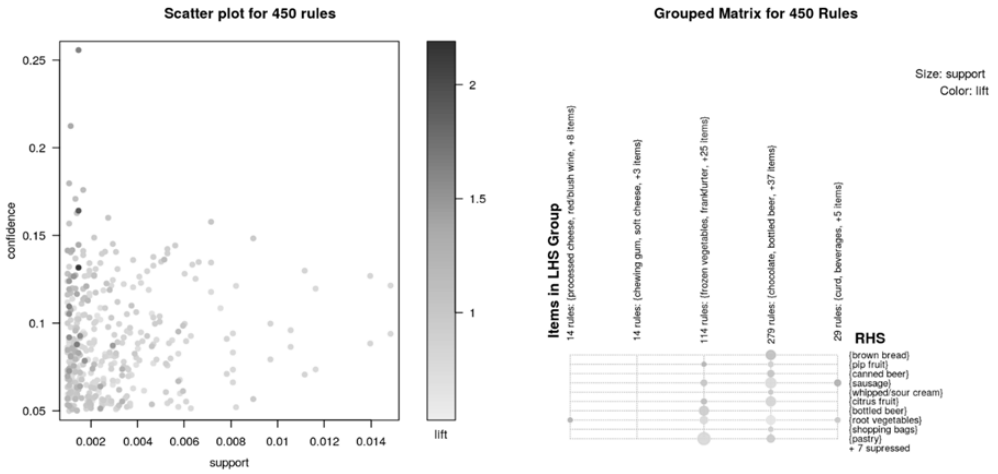

Figure 2: Visualization of the data mined association rule.
A) scatter plot; B) grouped matrix

Moreover, we visualize the 50 mined association rules using a graph representation. The graph provides a visual and interconnected view of the associations, allowing for a comprehensive understanding of the relationships and dependencies between different items or entities. Figure 3 illustrates a graph visualization of the 50 association rules mined by our model. Each node in the graph represents an item or entity, while the edges between the nodes depict the associations between them. The thickness or color of the edges may indicate the strength or confidence of the associations.

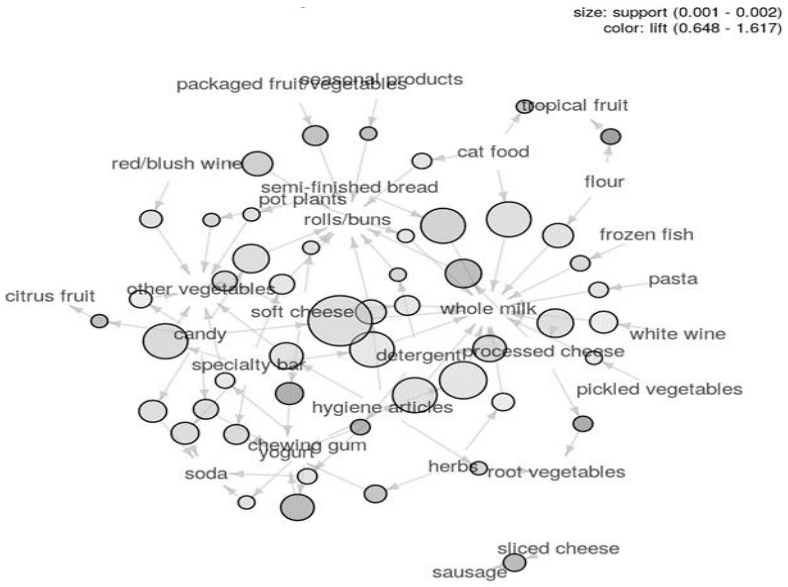

Figure 3: visualization of rules graph from our model.

5. Conclusion and Future Directions

In this paper, we proposed a methodology for mining semantic association rules from RDF data. We highlighted the significance of RDF data in representing structured information and discussed the motivation behind mining semantic associations. By employing the Apriori algorithm, we successfully extracted association rules and presented the results in terms of confidence, support, and lift. Our findings demonstrate the effectiveness of our model in discovering meaningful associations from RDF

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

data. The mined association rules uncover hidden relationships and patterns, enhancing our understanding of the underlying semantics in the data.

While our study presents promising results in mining semantic association rules from RDF data, there are several avenues for future research and improvement. Here are some potential directions to consider:

- Incorporating domain-specific knowledge: Enhancing the mining process by incorporating domain-specific knowledge, ontologies, or background knowledge can improve the quality and relevance of the discovered associations.

- Handling large-scale RDF datasets: Investigating scalable techniques and algorithms to handle large-scale RDF datasets is essential for mining associations in real-world scenarios with massive amounts of data.

- Considering temporal associations: Expanding the model to incorporate temporal aspects and analyzing associations over time can provide insights into evolving relationships and dynamics in the data.

- Utilizing advanced machine learning techniques: Exploring advanced machine learning techniques, such as deep learning or ensemble methods, can enhance the mining process and enable the discovery of more complex associations in RDF data.

- Evaluating the practical implications: Conducting empirical studies or case studies to evaluate the practical implications and effectiveness of the discovered associations in real-world applications or decision-making processes.

Funding: "This research received no external funding"

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

References

[1] Asadifar, Somayyeh, and Mohsen Kahani, Semantic association rule mining: a new approach for stock market prediction. 2017 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), IEEE, 2017.
[2] Barati Molood, Quan Bai, and Qing Liu, Mining semantic association rules from RDF data. Knowledge-Based Systems, 133, 183-196, 2017.
[3] Barati, Molood, Quan Bai, and Qing Liu, SWARM: An approach for mining semantic association rules from semantic web data. PRICAI 2016: Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, August 22-26, 2016.
[4] Heba R. Abdelhady, Mahmoud M. Ismail, Cardiovascular Diseases Forecasting using Machine Learning Models. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2) , 56-62, 2022.
[5] Alfrjani, Rowida, Taha Osman, and Georgina Cosma, A new approach to ontology-based semantic modelling for opinion mining. 2016 UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim), IEEE, 2016.
[6] LePendu, Paea, et al., Ontology database: A new method for semantic modeling and an application to brainwave data. Scientific and Statistical Database Management: 20th International Conference, SSDBM 2008, Hong Kong, China, July 9-11, 2008 Proceedings 20. Springer Berlin Heidelberg, 2008.
[7] Mohamed Saber, Efficient phase recovery system, IJEECS, 5(1), 2017.
[8] Alber S. Aziz, ,Hoda K. Mohamed, Ahmed Abdelhafeez, Unveiling the Power of Convolutional Networks: Applied Computational Intelligence for Arrhythmia Detection from ECG Signals. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2), 63-72, 2022.

Doi: https://doi.org/10.54216/JAIM.040105
Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023

[9] Douali, Nassim, et al., Case based fuzzy cognitive maps (CBFCM): new method for medical reasoning: comparison study between CBFCM/FCM. 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011). IEEE, 2011.
[10] Ismail Eyad Samara, Intelligent systems and AI techniques: Recent advances and Future directions. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2), 30-45, 2022.
[11] Mohamed Saber, A novel design and implementation of FBMC transceiver for low power applications. IJEEI, 8(1), 83-93, 2020.
[12] Sirichanya, Chanmee, and Kesorn Kraisak. Semantic data mining in the information age: A systematic review. International Journal of Intelligent Systems 36(8), 3880-3916, 2021.
[13] Abdulla Alsharhan, Natural Language Generation and Creative Writing A Systematic Review. Journal of International Journal of Advances in Applied Computational Intelligence, 1(1), 69-90, 2022.
[14] Hani D. Hejazi, Ahmed A. Khamee, Employees Motivational Factors toward Knowledge Sharing: A Systematic Review. Journal of International Journal of Advances in Applied Computational Intelligence, 1(1), 45-68, 2022.

Doi: https://doi.org/10.54216/JAIM.040105

Received: October 19, 2022 Revised: February 12, 2023 Accepted: June 17, 2023