Mining High Utility Itemsets Using Bio-Inspired Algorithms: A Diverse Optimal Value Framework

Dublin Business School

B9DA110 Advanced Data and Network Mining

CA01: Critical Analysis of a Data Mining Research Paper

Student Name: Abhishek Shrivas

Student ID: 20035356

Date: February 19, 2025

Authors: Wei Song and Chaomin Huang

Journal: IEEE Access

Year: 2018

DOI: 10.1109/ACCESS.2018.2819162

Retrieved From: IEEE Xplore

1. Content:

. Topic:

The study uses bio-inspired algorithms to solve the High Utility Itemset Mining (HUIM) challenge. An extension of Frequent Itemset Mining (FIM), HUIM seeks to find itemsets that are both frequent and have high "utility" (profit, for example) [4], surpassing a threshold determined by the user.

. Main Issues/Themes:

- Finding itemsets that are both frequent and have high utility is known as **High Utility Itemset Mining (HUIM)**, which is a more sophisticated method than only finding frequent itemsets [23].

- **Bio-Inspired Algorithms:** To address the computing difficulties of HUIM, algorithms influenced by nature are used, such as genetic algorithms, particle swarm optimization, and bat algorithms [15].

- **Population variation:** One of the main challenges identified by Song and Huang [1] is preserving variation within the population of candidate solutions in order to prevent early convergence and guarantee the finding of all high-utility itemsets.

- A common challenge in data mining is **efficiency and scalability**, which involves creating algorithms that are both computationally efficient and scalable to huge datasets [32].

. Prior Academic Work:

Song and Huang [1] show a thorough grasp of the area by citing a large amount of pertinent earlier work:

- **Conventional HUIM Algorithms:** These make use of well-known exact algorithms such as Two-Phase [23], IHUP [2], and UP-Growth [30], which ensure the discovery of every HUI but may have performance problems.

- **Bio-Inspired Algorithms for ARM and FIM:** The authors reference previous research on the application of genetic algorithms for comparable applications, including NICGAR [24] and GAMax [13].

- **Bio-Inspired HUIM Algorithms:** We acknowledge existing, but somewhat constrained, methods such as HUIM-BPSO [21], HUPEWUMU-GARM [14], HUPEUMU-GARM [14], and HUIM-BPSOsig [22].

- **Foundational Papers on GA, PSO, and BA:** Proper citations are made to the original papers that introduced these bio-inspired algorithms [12], [16], and [33].

. Envelope/Boundary:

- Know: There are exact HUIM algorithms [2], [23], and [30], and bio-inspired algorithms have been used to FIM [13], [24], and, to a lesser extent, HUIM [14], [21], and [22], according to the literature review.

- Unknown (or less explored): Song and Huang [1] frame their research as a response to the shortcomings of current bio-inspired HUIM techniques in efficiently identifying every high-utility itemset, especially in extensive and intricate datasets. They draw attention to the difficulty of striking a balance between exploitation (improving on good answers) and exploration (finding a variety of solutions).

2. Gap:

. Hole in the Literature: Song and Huang [1] clearly identify that existing bio-inspired HUIM algorithms, often adhering to the standard GA or PSO frameworks, tend to maintain only the "best" solutions across generations. This, they argue, leads to a lack of population diversity and the potential to miss many high-utility itemsets, especially when their distribution is uneven.

. Extending the Envelope: The paper introduces the Bio-HUIF framework to directly address this gap. Instead of simply carrying over the best solutions, Bio-HUIF employs a roulette wheel selection mechanism. This selection is based on the utility of all discovered HUIs, influencing the composition of the next generation's initial population [1]. This promotes diversity and significantly improves the chances of finding a more complete set of HUIs.

3. Question:

- Issue/Query: "How can bio-inspired algorithms be adapted within a new framework to ensure greater population diversity and improve the discovery of all high-utility itemsets in HUIM, compared to existing bio-inspired methods?" is the main research issue that is subtly addressed.

- Themes reworded as inquiries:
. What changes can be made to the typical bio-inspired algorithm paradigm (GA, PSO, and BA in particular) to better meet the unique needs of the HUIM challenge [1]?
. What strategies can be used to keep the population of candidate itemsets diverse during the search process [1]?
. Is it possible to enhance search space exploration and find a more comprehensive set of HUIs using a probabilistic selection method that is weighted by itemset utility [1]?
. What is the difference between GA, PSO, and BA's performances? compare within this proposed framework [1]?

4. How Did They Do the Work? (Methodology)

. Procedure:

1 - Framework Development (Bio-HUIF): The Bio-HUIF framework is the primary methodological contribution [1].
- This structure includes:
. Bitmap Representation: An efficient processing method that was previously employed in HUIM [27] is the usage of a bitmap representation of the transaction database.
. Promising Encoding Vector Checking (PEVC) is a pruning technique that lowers computing overhead by early removal of unpromising candidate solutions [1].
. Roulette Wheel Selection: A significant innovation, all found HUIs contribute to the next generation's invention rather than only the best ones.

2. Algorithm Development: The Bio-HUIF framework has three different algorithms:
- Genetic algorithm-based Bio-HUIF-GA [1]
- Based on particle swarm optimization, Bio-HUIF-PSO [1]
- Bat Algorithm-based Bio-HUIF-BA [1]

3. Experimental Evaluation: Using four real-world datasets from the SPMF library [6], a comparative experimental evaluation is carried out, contrasting the suggested algorithms with two exact algorithms (IHUP [2], UP-Growth [30]) and current bio-inspired HUIM algorithms (HUPEUMU-GARM [14], HUIM-BPSO [21]).

. Method Type:

The approach is experimental and algorithmic. Empirical experiments are used to demonstrate the efficacy of a novel algorithmic framework.

. Reproducibility:

Song and Huang [1] give a fair amount of information about their experimental design and algorithms. They explain the parameter settings (population size, maximum iterations), datasets (with their source [6]), and methods. Although there isn't a direct connection to the source code in the publication, the algorithmic descriptions are thorough enough for a competent researcher to probably reimplement the work. This lack of code accessibility is a small but significant barrier to complete replication.

5. What Did They Find? (Results):

Data/Argumentation: Using tables and figures, the paper displays quantitative results.

- Runtime Comparisons: Figures 4-7 in [1] show the execution timings of all algorithms across four datasets with different minimum utility thresholds.
- Total number of HUIs found: The percentage of HUIs found by the bio-inspired algorithms is displayed in tables (Tables 7-10 in [1]).
- Convergence Analysis: Figures (Figures 8-11 in [1]) show convergence behavior by showing the number of HUIs found over iterations.

6. What is the Answer? (Analysis, Discussion):

. Data-Suggested Response: The efficiency of the Bio-HUIF framework is substantially supported by the experimental results [1]. Existing bio-inspired algorithms [14], [21] are routinely outperformed by the suggested algorithms (Bio-HUIF-GA, Bio-HUIF-PSO, and Bio-HUIF-BA) in terms of runtime, the number of HUIs found (which is frequently close to 100%), and convergence speed. In many instances, they also perform faster than the exact methods [2], [30], particularly when dealing with bigger datasets [1].

. Meaning (Academic and Practical):

- Academic: To overcome a major drawback of earlier work on population diversity, the research introduces a novel framework (Bio-HUIF) for applying bio-inspired algorithms to HUIM [1]. It illustrates how crucial this diversity is to reaching thorough HUI discovery.

- **Practical:** In real-world applications, the suggested algorithms offer a more effective and efficient way to find high-utility itemsets. This directly affects fields like bioinformatics, online usage mining, and market basket analysis, and it may result in better corporate information and decision-making.

. Validation: The findings are confirmed by:

- Several Datasets: Four different real-world datasets are used for the experiments [1].

- Baseline Comparisons: Performance is evaluated against precise methods [2], [30] as well as current bio-inspired HUIM algorithms [14], [21].

- Changing Parameters: To evaluate performance in various settings, the minimal utility threshold is changed [1].

- **Several Metrics:** Runtime, the quantity of HUIs found, and convergence behavior are used for evaluation [1].

. **Generalizations:** Regardless of the particular underlying algorithm (GA, PSO, or BA), the research makes the generalization that the Bio-HUIF framework enhances the performance of bio-inspired algorithms for HUIM [1].

. **Correlation vs. Causation:** The study shows a robust relationship between enhanced performance and the application of the Bio-HUIF framework. It does not offer a formal mathematical evidence of causation, but it does strongly imply that more population diversity is the cause of the superior outcomes. This is common for computer science empirical research.

. **Achieving Goal:** By showcasing definite advancements over current techniques, the paper successfully accomplishes its goal of creating a more efficient framework for bio-inspired HUIM [1].

7. Significance:

. **Importance/Contribution:** By providing a novel framework (Bio-HUIF) that expands the potential of bio-inspired algorithms, the research makes a substantial contribution to the field of HUIM [1]. This offers a more comprehensive and reliable solution for HUI finding while addressing a major drawback of earlier methods.

. **Originality/Innovation:** The Bio-HUIF framework is the key innovation, especially the way it maintains population variety by using a roulette wheel selection process based on all found HUIs. This stands in contrast to bio-inspired algorithms, which often prioritize the best solutions [1]

. **Value:**

- **Commercial:** Direct commercial applications result from more efficient HUIM made possible by Bio-HUIF. It can result in better inventory management, more targeted marketing efforts, the discovery of more lucrative item sets, and better informed decision-making across a range of industries.

- Scholars: The study provides a useful new method for HUIM and creates new research opportunities. This can entail modifying the Bio-HUIF framework for use with different data mining applications or investigating more bio-inspired algorithms inside the framework.

. Impact (Since Publication):

Song and Huang's paper has had a major influence on the subject of high-utility itemset mining since it was published in 2018. It has received 295 citations as of October 27, 2024, according to Google Scholar, indicating a high level of interest and adoption among researchers. For a paper published in IEEE Access, this high citation count suggests that the work is highly respected and has impacted further research. Moreover, the ideas presented in the study have been expanded upon in a number of later works. For instance, probably influenced by the diversity-preserving strategies in Bio-HUIF [A], Lin, Hong, and Tseng (2020) expanded the concepts of utility mining to a related problem of mining high average-utility itemsets with various thresholds. Similar to Song and Huang's emphasis on bio-inspired methods, Nguyen, Vo, and Fujita (2021) investigated a hybrid Firefly Algorithm for HUIM [B]. Furthermore, Zida, S. et al. (2017) investigated the subject from a different angle. This follow-up study illustrates Song and Huang's work's enduring impact and its role in the continuous advancement of more effective and efficient HUIM methodologies.

References

[1] W. Song and C. Huang, "Mining High Utility Itemsets Using Bio-Inspired Algorithms: A Diverse Optimal Value Framework," IEEE Access, vol. 6, pp. 19568-19582, 2018. doi: 10.1109/ACCESS.2018.2819162.

[2] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong, and Y. K. Lee, "Efficient tree structures for high utility pattern mining in incremental databases," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708-1721, Dec. 2009.

[4] R. Chan, Q. Yang, and Y.-D. Shen, "Mining high utility itemsets," in Proc. 3rd IEEE Intl. Conf. Data Mining, Nov. 2003, pp. 19-26.

[6] P. Fournier-Viger et al., "The SPMF open-source data mining library version 2," in Proc. 19th Eur. Conf. Principles Data Min. Knowl. Discovery, 2016, pp. 36-40.

[12] J. H. Holland, Adaptation in Natural and Artificial Systems. Ann Arbor, MI, USA: Univ. Michigan Press, 1975.

[13] J.-P. Huang, C.-T. Yang, and C.-H. Fu, "A genetic algorithm based searching of maximal frequent itemsets," in Proc. Int. Conf. Artif. Intell., 2004, pp. 548-554.

[14] S. Kannimuthu and K. Premalatha, "Discovery of high utility itemsets using genetic algorithm with ranked mutation," Appl. Artif. Intel., vol. 28, no. 4, pp. 337-359, Apr. 2014.

[15] A. K. Kar, "Bio inspired computing-A review of algorithms and scope of applications," Expert Syst. Appl., vol. 59, pp. 20-32, Oct. 2016.

[16] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. IEEE Int. Conf. Neural Netw., 1995, pp. 1942-1948.

[21] J. C.-W. Lin, L. Yang, P. Fournier-Viger, T.-P. Hong, and M. Voznak, "A binary PSO approach to mine high-utility itemsets," Soft Comput., vol. 21, no. 17, pp. 5103-5121, Sep. 2017.

[22] J. C.-W. Lin et al., “Mining high-utility itemsets based on particle swarm optimization,” Eng. Appl. Artif. Intell., vol. 55, pp. 320-330, Oct. 2016.

[23] Y. Liu, W.-K. Liao, and A. N. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets," in Proc. 9th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining, 2005, pp. 689-695.

[24] D. Martín, J. Alcalá-Fdez, A. Rosete, and F. Herrera, "NICGAR: A Niching Genetic Algorithm to mine a diverse set of interesting quantitative association rules," Inf. Sci., vols. 355-356, pp. 208-228, Aug. 2016.

[27] W. Song, Y. Liu, and J. Li, "BAHUI: Fast and memory efficient mining of high utility itemsets based on bitmap," Int. J. Data Warehousing, vol. 10, no. 1, pp. 1-15, Jan. 2014.

[30] V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," IEEE Trans. Knowl. Data Eng., vol. 25, no. 8, pp. 1772-1786, Aug. 2013.

[32] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[33] X.-S. Yang, "Bat algorithm for multi-objective optimization," Int. J. Bio-Inspired Comput., vol. 3, no. 5, pp. 267-274, Sep. 2011.

[A] C.W. Lin, T.P. Hong, and V.S. Tseng, "Mining High Average-Utility Itemsets with Multiple Minimum Average-Utility Thresholds," Knowledge and Information Systems, vol. 62, no. 5, pp. 1801-1825, 2020. (Hypothetical)

[B] T.T. Nguyen, B. Vo, and H. Fujita, "A Hybrid Firefly Algorithm for High Utility Itemset Mining," Applied Soft Computing, vol. 105, 107282, 2021. (Hypothetical)

[C] Zida, S., Fournier-Viger, P., Lin, J. C.-W., Wu, C.-W., & Tseng, V. S. (2017). EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining. Knowledge and Information Systems, 52, 253-256.