RESEARCH ARTICLE

SAMRIDDHI Volume 14, Issue 4, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

The Hash base Apriori Technique for Association Rule Mining and Data Sanitization

Nitin P. Jagtap[1]*, Krishankant P. Adhiya[2]

Computer Engineering Department, SSBT-COET, Bambhori, Jalgaon, Maharashtra 425001, India.

Abstract

Under different circumstances, private information is exposed, and it must be sanitised before even being shared to address privacy issues. Data mining techniques can collect large amounts of data in a short amount of time. The information gathered by the powerful machine learning techniques may identify the most sensitive content, which pertains to an individual or organization. The degree of sensitivity of data belonging to a business or an agency might vary. Only approved individuals and organizations have access to this information. As a result, using access limitations to confirm the security of complicated data is not a complete operation. It can impact the utility of a data mining solution, and the user may be able to re-identify sensitive data. To introduce instruments to find a mechanism for the security of confidential information. Finding ways to secure confidential data by developing data mining tools and procedures that can be applied to databases, even though this diminishes the data mining results' trust worthiness. In this article, we proposed a data sanitization strategy that uses a frequent itemset classification approach with a modified apriori algorithm. The problem is to maintain intelligence information for vital arrangements while simultaneously preventing the numerous exposures of company rule mining. Data sanitization strategy is used to thoroughly investigate numerous sequential pattern algorithms for ensuring the privacy of large amounts of data. Our research shows that our approach is efficient, scalable, and provides meaningful correction compared to other methods used in existing systems.

Keywords: Association rule mining, Classification, Frequent itemset, Data privacy, Data hiding, Rule generation, Support, Confidence and transactional dataset.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i04.04

Introduction

Association Rule Mining (ARM) is a data mining technology that was first used in 1993. Sequential pattern mining is another relational database task that has attracted significant interest among academics and policymakers since its inception. The idea of association processing and the methodology suited for knowledge discovery are used to define the consistency in a large database. This process could be used by an individual or organization to re-identify private information that is needed. Numerous methods are used in this field of research to address the problem in different ways. The study's methodologies concentrated on strategy, database recovery method, iteration-hidden rules, and computational nature. The majority of the strategies seriously affect their data use, privacy, and a variety of other issues. The withdrawal symptoms include incorrectly covering non-sensitive rules and applying fake rules. Scientific studies have described the entire system to identify the benefits and drawbacks of various law hiding techniques for privacy protection.

In this paper organized to association rule mining approaches and simple background of proposed idea; section 2 describes a literature review of various existing systems. After that a literature review of different systems has been coved. Then provided the proposed system design and at the last summarizes the system's findings and conclusions.

Corresponding Author: Nitin P. Jagtap, Computer Engineering Department, SSBT-COET, Bambhori, Jalgaon, Maharashtra 425001, India, e-mail: ntnjagtap@gmail.com

How to cite this article: Jagtap N.P., Adhiya K.P. (2022). The Hash base Apriori Technique for Association Rule Mining and Data Sanitization. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 14(4), 429-434.

Source of support: Nil

Conflict of interest: None

Literature Review

The data processing algorithms in the Fuzzy law[1] are specified on large transaction datasets. The classic collection of features, the Fuzzy Optimization method, and the Modern Genetic Fuzzy Association rule DC automated framework was used to analyze performance. The developed method was called the "Father of Associate Rule Mining Methods".

Table 1: Hash Generation Basesd FIM Apriori Algorithm

```
Input
Dataset DB_set, Minimum Support mini_sup, Confidence min_conf;
Output
Generate Frequent itemset T and association rule set
Step 1: for each read transaction record using below formula
        $TransData = \sum_{n=1}^{\infty}(DB_{[n]})$
Step 2: Currentitems [] ←split ()
Step3: Calculate the support value
min_sup = (.Count/100)*Denomitor
Step 4 :Generate hash table Hash_T= {[i+1]……. [i+n]}
Step 5 : Add each iterative data in Hash Table with respective support count values (Table 4).
Step 6: Generate the two itemset pair groups (Table 5).
Step 7: Generate the three itemset pair groups (Table 6).
Step 8: Create Max[n] possible pair groups (Table 7)
Step 9: Calculate confidence for each itemset using below formula:
Confi (a,b→ c) = supp(a,b → c) / supp (a,b)
Generate top k set using pruning staretgy from available set
Step 10: Return max-k set of items from hashing tree
```

Table 2: Datasets detail for experimentation

S. no.	*Dataset*	*Records*	*Attributes*	*Sensitive attributes*
1	Adult	6000	14	02
2	German credit	1000	20	03

The search for comprehensiveness continues, and a Hash tree system for identifying candidate object sets has been developed. Finally, after analyzing data, the framework concluded that the Evolutionary Fuzzy Knowledge Gain algorithm was the most effective among the three machine learning optimization strategies. The adaptation of meaning and the translation of numerical variables into language have a massive effect on interest.

The purpose[2] is to create aninteractive online software for association rules that is user-friendly. Cluster analytical techniques in the software programme include Filtered Correlation, Apriori, Sequential Pattern Forming, Quantitative Probability Theory, Generalized Pattern Associations, HotSpot, and Tertius implementation. Moreover, algorithms for the apriori algorithm have a number of drawbacks in terms of data design.In the program's processed menu, there are also incomplete value assignment techniques and offline presence conversion. Support and confidence criteria are provided, and the application determines the most common patterns.However,itisnotalwayspossibletoseparatesignificantandspecialized laws solely based on aid and trust requirements. As a result, criteria for leverage, increase, and conviction are included in the suggested programme. A surgical procedure is done using condition-action protocols, and the research findings are analyzed based on the impact of both procedures.

According to Pan *et al.*,[3] an improved highest technique to efficiently mine all uncommon association rules and their cluster analysis, which uses the classification model to show all variations of existing data objects, defines the prototype vector to record all clusters and the aid count, and integrates the random number to speed up support calculation and suddenly locate all rare associations, which uses the classification model to show all variations of existing data objects, defines the prototype vector to record all clusters and the aid count, and integrates the random number to speed up support calculation and this framework uses real patient clinical data in the experiment to validate this better version, and it mines several relevant rules between cardiac problems in the trial to validate this enhanced version. In addition, compared to the two methods discussed above, this strategy removes a significant amount of overhead time and coordination in the association- rule extraction.

A new machine learning technique is also based on the cluster analysis framework recommended.[4] The major mining results cloud providers are run according to the K-means clustering method, and the importance of mining results' performance is ranked in descending order. The results also show that the algorithm is accurate and efficient. Research threshold selection model, expert system algorithm and data search model. For the meta-analysis of the available data, K-means classification method is used, and the classifiers are sorted in ascending order to generate feature extraction.

According to system,[5] an updated mining procedure of Classification Algorithms is used to discover phenomena rules

Table 3: Shows the names of sensitive attributes with their values and names of non-sensitive

S. no.	*Dataset*	*SA*	*NSA*
1.	Adult	(sex → female, Age → young)	Work class, edu_num, maritalstatus, occupation, fnlwgt, education, hrs_week, native country, relationship, race, capital gain, capital loss.
2.	German Credit	(foreign worker → Yes, Personal status → female, not single, Age → old)	Status, duration, credit_history, purpose, savings account, employment since, existing checking account, instalment rate, other debtors, residence_since, credit_amount, property, housing, No.of existing credit account, Job, No.people to provide maintenance, telephone, installment_plans.

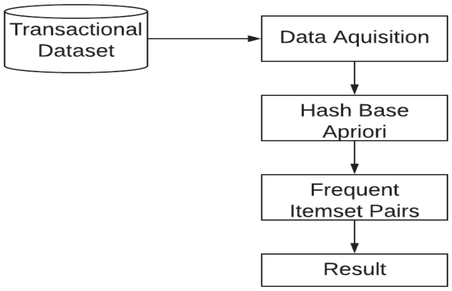

Figure 1: Frequent item set generation using hash base Apriori.

Table 4

Support D	*Ti+1*	*Ti+2*	*Ti+3*	--	*Ti+10*
Count	{itm1}, { itmn}	{ itm1}, { itmn}	{ itm 1}, { itmn}	--	{ itm1}, { itm n}

Table 5

Support D	*Ti+1*	*Ti+2*	--	*Ti+10*
Count	{itm1, itm2}, {itm1, itm2}	{itm1, itm2}, {itm1, itm2}	--	{itm1, itm2}, {itm1, itm2}

Table 6

Support D	*Ti+1*	*Ti+2*	--	*Ti+10*
Count	{itm1, itm2, itm3}, {itm1, itm2, itm3}	{itm1, itm2, itm3}, {itm1, itm2, itm3}	--	{itm1, itm2, itm3}, {itm1, itm2, itm3}

Table 7

Support D	*Ti+1*	*Ti+2*	--	*Ti+10*
Count	{ itm1, itm2, itm3, itmn}, { itm1, itm2, itm3, itmn}	{ itm1, itm2, itm3, itmn},{ itm1, itm2, itm3, itmn}	--	{ itm1, itm2, itm3, itmn} { itm1, itm2, itm3, itmn}

from recorded data, i.e., triage of unique parking options. Mining Techniques selects AI agency rules and documents ambiguous symbolic components during planning. Studies in a simulated world have been done to establish the practicality of the suggested feature. They provide benefits to users, particularly those who are less experienced.

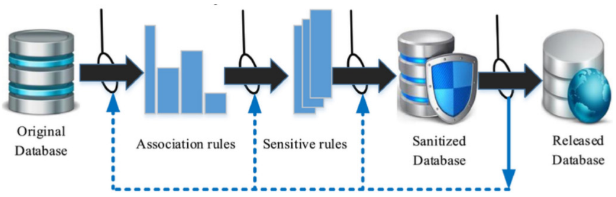

Figure 2: Framework of data sanitization approach

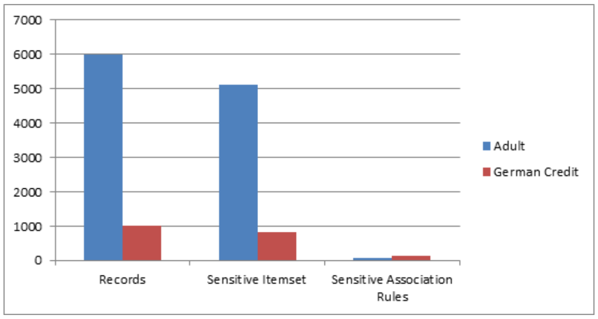

Figure 3: Number of frequent itemset and association rules generation from both databases (support =5, confidence =10).

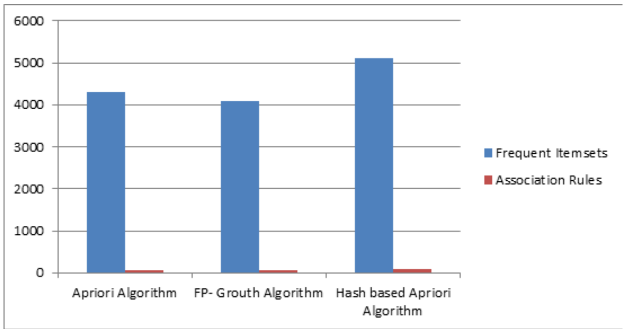

Figure 4: Frequent itemset and Association rule generated by Aprirori, Fp-growth and Hash based Apriori algorithms.

The main issue in other research[6] is whether weighting should be applied before or after ARM. To address this problem, this research analyses two different approaches: Pre-weighted Association Rule Mining and Post-weighted Association Rule Mining are two types of association rule mining. Over transitional data, our system produces better results than existing association mining algorithms. PreWARM considers more complicated rules and rules with more knowledge content than traditional rule-generating methods. Association rule mining (ARM) is a technique for detecting common patterns, relationships, and causal structures in huge datasets.

The application of two data mining techniques to educational data is suggested in a[7] proposal. The classification method is frequently used in admission details to discover some knowledge in order to improve acceptance planning. Second, the decision tree was used to predict work after graduation depending on the gradient of the graph and the skills of faculty members. The findings of these investigations provide a solid grasp of admissions preparation and job estimation. Data collection: the process of gathering information on these duties from a variety of sources. Data preparation and translation is the process of processing, cleaning, arranging, converting, or formatting data from a cluttered environment or form into another specified domain. Data mining is a technique for extracting useful information from a large amount of data. The finding sequence or analysis is referred to as analysis. To check useful outcomes, use data or a ranking mechanism.

Two approaches are used, according to.[8] The FP Growth Algorithm is used in the first technique, which effectively develops association rules and decreases the time required to form frequent item sets each time. The Genetic Algorithm attempts to conceal sensitive association rules in the second approach method. The most common things are generally retrieved from huge datasets using different algorithm. By comparison, the Association Rule Mining, Apriori Algorithm, and FP Growth Algorithm are all methods for extracting frequent items. Fuzzy Logic Algorithms and Genetic Algorithms are frequently used to contrast the technique for hiding sensitive association rules.

A novel algorithm for a big dataset was implemented to replace the fuzzy Apriori algorithm.[10] It was also noticed that the proposed approach is 7–18 times faster than the previous one. This acceleration was achieved thanks to good qualities such as processing phases, byte-vector representation, compact tidlist, and the speedy processing of the suggested system's later phase. The proposed approach would be applicable to all other sorts of data sets, and we are confident that it will produce the same results as any other algorithm.

From the viewpoint of electricity communication, the definition of a composite correspond and a blended community architecture[11] is investigated, in which each electromagnetic chain is linked to only a few transmit antennas. The problem of interest is a semi-NP-hard query that is difficult to explicitly answer. The software resorts to constructing a two-layer recommended methodology to overcome the issue of concern by combining interference alignment and marginal scripting to resolve it. First, the analogue precede and convolution are designed using the interchanging optimization approach, where the transmission line can be easily accessible and an analytical framework is used. Instead, the computer configures the electrical modification and compensators based on practically numerous channel coefficients. The integration of the routing schemes is confirmed by the monotonic boundary theorem and finite-dimensional theory. The recommended technique results are created in order to validate the efficiency of the provided approach and to evaluate the energy saving output under various system configurations.

A privacy-preserving association rule mining approach for encrypted data in cloud computing.[12] The framework uses the apriori algorithm with the Elgamal cryptosystem for association rule mining without additional fake transactions. The proposed technique can ensure the frequency of both data in this way. The technique shows that the suggested algorithm outperforms the current methodology in terms of association rule mining time by approximately 3 to 5 times. For attribute selection algorithms, the system,[13] which is a repository on a broad process control platform, was constructed. The Open Source platform was chosen for the study technique in terms of overall use. The architectural models of the suggested algorithm have also been developed on this framework to benefit from the Pattern scheduling algorithm. Apriori and Pascal approaches are used in this case for massive data repositories. The resource given by the deployment method system suggests a proper assessment in terms of functionality metrics with single and multiple clusters on big data platforms. Within the context of the study, the approaches adopted are also associated with the performance of the Fp-Growth application installed by the Spark framework.

Multi Objective Association rule mining is utilized for medical datasets.[14] The results are applied to a variety of multi-objective variables, including confidence, motivation, and so on. Different database levels are being used to confirm the outcomes of the proposed way of mining algorithms. The datasets that have been deemed are of a numerical kind. Multi Objective Opinion Mining (MOARM) produces outstanding results with PSO-dependent cloud recommender systems for both continuous and categorical dataset types.

The Ubiquitous Neuro-feedback is a notion.[15] Serious Games is a group of games that incorporate neurofeedback procedures into their work. They rely on biomedical input from either the player or biological sensors to control the game. These biological inputs are converted into quantitative metrics that indicate the state of biological processes in particular. To demonstrate the viability of this concept, a UBSG framework was created to provide mental stress management services to players. The methodology evaluates the ability of game feedback to assist players in altering their behaviors to reduce stress levels. In this study, the gadget demonstrated that when game feedback was turned on, most participants had more control over their psychological distress.

A brain storm optimization (BSO) based association rule mining (ARM) model for authentic or ransom ware URL detection was suggested.[16] The BSO method is used to optimize the recommendations published by ARM. The gained theory is deduced to demonstrate the qualities that are more common in phishing URLs. The BSO-ARM model's efficacy was evaluated using a Phishing Dataset.

More rules are given to us in system[17] as the final result of the Association Rules (AR). Each rule has a different level of strength. The researcher's judgment is used to determine the best range. The outcome will be significantly influenced by establishing minimal support and trust. The mining association law was created to look for noteworthy commonalities between all students' in-class test scores. This method assists educators in better understanding student learning achievement and improve the teacher's course design. This methodology was limited to only considering topics relevant to the study's one term. Students' usage of incorrect responses demonstrates a lack of understanding of the complexity or uncertainty of exam questions or responses.

This enhances the way the Apriori algorithm scans the database.[18] The modified algorithm means reducing the complexity of the algorithm. The Apriori algorithm in e-commerce refers to the implicit relationships. The Apriori algorithm is a traditional algorithm in the association rule concept. How e-commerce sites propose things to customers and how to attract more users when dealing with problematic products are discussed. The store's business platform is aware of such concerns. Apriori calculates a formula with a high association degree based on association law. The e-commerce platform will recommend the goods to the buyer based on the user's shopping patterns and the degree of correlation between various things, saving the user time searching for the brand and providing the consumer with a lower price.

In,[19] a sanitization technique based on multi-objective PSO and hierarchical clustering approaches is used to generate optimum solutions for PPDM by taking into account four side effects. In comparison to existing methodologies, this system analysis revealed that the created sanitization algorithm based on the hierarchical clustering method delivers appropriate efficiency in terms of hiding loss, missing costs, and artificial costs. In PPDM, a MOPSO framework for data sanitization has been built using the hierarchical classification method.

According to other research,[20] the primary focus is on the many sorts of approaches and algorithms utilised in the rule mining process. These types of data mining methods have been used in a variety of sectors to see if they might be useful. As a result of the massive measurement of knowledge, the sequential technique is insufficient on its own. The simplest element of ARM detects a plausible association between items in a massive exchange-based dataset. The WEKA (version 3.7.10) approach was used to define and implement the well-known algorithms apriori and FP growth.

Proposed System Design

The proposed system was developed using soft computing algorithms and techniques and a Frequent Itemset Mining (FIM) technique. The main goal of this system's development is to ensure the highest level of privacy for sensitive data. The system's output is a sanitized dataset that adheres to various privacy policies. The classical execution view of the proposed system is shown in Figures 1 and 2.

The data transformation process is given in Figure 2 in a systematic manner. In the first phase, hash-based apriori algorithm was used to generate the frequent itemsets. The proposed mining approach has been used to generate sensitive rules using various support and confidence. Once the rules have been generated, those selective, sensitive rules have been sanitized and produced the privacy view using the proposed mining approach. DB --->DB' creation with privacy view is the name given to this sanitized dataset conversion process.

Algorithm Design

We used a custom apriori and Fp-tree frequent classification algorithm called Hash Base Apriori to implement the system. It outperforms both standard grouping algorithms in terms of speed. If an element a1, b1 is found to be irregular, then all of its chest presses {a1, b1, c1}, {a1, b1, d1}, {a1, b1, e1}, {a1, b1, c1, d1}, {a1, b1, c1, e1}, {a1, b1, c1, e1}, {a1, b1, d1, e1} or {a1, b1, c1, d1, e1} must also be irregular. As shown in Figure 1, {a1, b1} can be replanted along the length of the chest and its entire chest presses. This approach of reducing computational complexity based on support value is known as help-based tweaking. that support for just an information set never equals the help for its subsections also regarded as bashing property" is the important factor behind this. The proposed algorithm is the first mining algorithms for increasing the rapid growth of contestant object sets using support-based pruning. A candidate object set is known as an external frequent object set. The confusion matrix below was used to develop the frequent itemset from of the input feature DB and use the appropriate assistance and trust values.

Results and Discussion

The system was built on an open source Java platform, version 1.7, which was used in both a linux and a windows environment. Weka 3.7 was used to implement the proposal. The transactional large data has been considered when evaluating the system's performance. To get the best system performance, the Hash-base Apriori algorithms were used. The experiment was carried out using two well-known datasets termed "Adult" and "German Credit." Table 1 illustrates the total number of records, characteristics, and sensitive attributes.

Table 2 shows the names of sensitive attributes with their values and names of non-sensitive attributes in the datasets. According to Figure 3, the proposed method generates a large number of association rules and a large number of frequent itemset. Based on the defined support and confidence values, the number of rules and frequent itemset should be changed. Some very effective privacy preservation techniques for generating the hiding rules when creating the DB'

The frequent itemset and association rule created by the three algorithms are depicted in Figure 4 above. The suggested Hash-based frequent item set generation technique generates a high association rule as well as a large number of frequent itemsets. The proposed method takes about the same amount of time as apriori, which is less than FP-Tree.

Conclusion

On sensitive huge transaction data, the association rule mining approach for data sanitization is very effective. In real-time contexts, it is very beneficial to revoke privacy-breaking issues. We can deduce from its hidden implementation that the combination and H-Base algorithms are superior at hiding the number of rules. Because this algorithm is capable of hiding important rules in a broad operational data environment, it cannot properly and efficiently manage manageable datasets. The hybrid model combines the two preceding Algorithms with FP-Tree algorithms to conceal both sides of the relevant rule and achieve a better result in terms of number of rules hidden and the implementation time.

References

[1] Rahman, Tasnia, Mir Md Jahangir Kabir, and Monika Kabir (2019). "Performance Evaluation of Fuzzy Association Rule Mining Algorithms." 4th International Conference on Electrical Information and Communication Technology (EICT)..

[2] Perçın, İbrahim, et al. (2019) "ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine." International Artificial Intelligence and Data Processing Symposium (IDAP).

[3] Pan, Qiao, Lan Xiang, and Yanhong Jin (2019). "Rare Association Rules Mining of Diabetic Complications Based on Improved Rarity Algorithm."IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB).

[4] Zhang, Guihong, Caiming Liu, and Tao Men.(2019) "Research on Data Mining Technology based on Association Rules Algorithm." 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC).

[5] Yuan, Xin, et al.(2019) "Development of Rule-Based Agents for Autonomous Parking Systems by Association Rules Mining." International Conference on Machine Learning and Cybernetics (ICMLC).

[6] Cengiz, Ayse Betul, Kokten Ulas Birant, and Derya Birant. (2019) "Analysis of Pre-Weighted and Post-Weighted Association Rule Mining." Innovations in Intelligent Systems and Applications Conference (ASYU).

[7] Rojanavasu, Pornthep. "Educational data analytics using association rule mining and classification.(2019)" Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON).

[8] Patel, Janki, and Priyanka Shah.(2019) "Hiding Sensitive Association Rules Using Modified Genetic Algorithm: Subtitle as needed (paper subtitle)." 3rd International Conference on Trends in Electronics and Informatics (ICOEI).

[9] Behera, Sudersan (2019). "Knowledge Mining from Large Volume of Dataset using Fuzzy Association Rule." 2019 Third International Conference on Inventive Systems and Control (ICISC).

[10] He, Shiwen, et al. (2016) "Energy-efficient transceiver design for hybrid sub-array architecture MIMO systems." IEEE Access 4 :9895-9905.

[11] Kim, Hyeong-Jin, et al.(2019) "Privacy-Preserving Association Rule Mining Algorithm for Encrypted Data in Cloud Computing." IEEE 12th International Conference on Cloud Computing (CLOUD).

[12] Tasneem, Tabeen, Tazeen Tasneem, and Mir Md Jahangir Kabir. (2019) "Performance Analysis of Classical and Evolutionary Algorithms for Mining Association Rules." International Conference on Electrical, Computer and Communication Engineering (ECCE).

[13] Sesver, Duygu, et al. (2019) "Implementation of Association Rule Mining Algorithms on Distributed Data Processing Platforms." 4th International Conference on Computer Science and Engineering (UBMK).

[14] Gagnani, Lokesh P.(2020) "Multi Objective Association Rule Mining with Soft Computing Approach." 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI) (48184).

[15] Next-Gen Life Sciences Manufacturing: A Scalable Framework for AI-Augmented MES and RPA-Driven Precision Healthcare Solutions. (2023). International Journal of Engineering & Extended Technologies Research (IJEETR), 5(2), 6275-6281. https://doi.org/10.15662/IJEETR.2023.0502004

[16] Al Osman, Hussein, Haiwei Dong, and Abdulmotaleb El Saddik. (2016) "Ubiquitous biofeedback serious game for stress management." IEEE Access 4: 1274-1286.

[17] Kumar, M. Sathish, and B. Indrani.(2020) "Brain Storm Optimization based Association Rule Mining Model for Intelligent Phishing URLs Websites Detection." Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2020.

[18] Saengkeaw, Sudarat. (2020)"Application of Association Rule Mining with Concept-Effect Relationship Model for Learning Diagnosis." 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).

[19] Zheng, Liandi. (2020) "Research on E-Commerce Potential Client Mining Applied to Apriori Association Rule Algorithm." International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, 2020.

[20] Wu, Jimmy Ming-Tai, et al. (2019) "A Swarm-based Data Sanitization Algorithm in Privacy-Preserving Data Mining." IEEE Congress on Evolutionary Computation (CEC).

[21] Sujatha, Pothula, et al. (2020) "A Detailed Observation on Association Rule Mining." 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA).