Journal of Information Security Research

DLINE JOURNALS

Print ISSN: 0976-4143
Online ISSN: 0976-4151

JISR 2025: 16 (1)
https://doi.org/10.6025/jisr/2025/16/1/1-8

Secured hash Techniques for the Analysis of Physical Health Based on Third Party and Apriori Algorithm

Nan Zhang
College of Physical Education and Health
Xinxiang Vocational and Technical College
Xinxiang, 453006, China
hnxxzn@163.com

ABSTRACT

This article aims to conduct empirical testing on 6043 students from 7 high schools to explore the data background and potential risks that third-party physical health measurements can reveal. Through Apriori Analysis of algorithms, we have concluded that there is a certain degree of similarity between the results and reference values of most indicators when measured by a third party. Therefore, this paper can effectively reveal the potential risks to effectively control and reduce the possible risks when preventing and treating diseases. Through the Apriori algorithm, we found that six rules with more than 10% support showed relatively small interaction between the Standing long jump and other sports. Therefore, we concluded that the athletes from seven middle schools had relatively good physical conditions, muscle elasticity, strength, and durability. In addition, data from third parties can more accurately reflect athletes' exercise status.

Keywords: Apriori Algorithm, Third-Party Testing, Student Physical Health

Received: 11 September 2024, Revised 9 December 2024, Accepted 20 December 2024

1. Introduction

With the development of technology, more and more data are being collected and processed, which can be deeply researched and utilized through data mining methods to obtain more practical value. Especially for university physical education courses, the Apriori algorithm can more accurately discover and reveal the mutual influence between course content, thereby better guiding and cultivating students' physical and mental health. The Apriori algorithm explores the interaction of various elements in the dataset by gradually increasing and decreasing difficulty, and it uses probability to describe the association of these elements [2]. However, this algorithm can also result in a higher computational burden, slower mining speed, and the possibility of many meaningless,

erroneous, or meaningless associations [3]. By using grey correlation analysis, we can accurately study the mutual influence and interaction of different factors in the evolution history of grey systems. Pizzulli V A proposed a new method, Xie H, in his in-depth physical fitness test data research. It uses the Apriori method to effectively explore the mutual influence between BMI and body fat rate in a layer-by-layer iterative manner, thus solving the problem of spearheading traditional obesity evaluation models. In addition, this method also has high support and confidence, making it one of the currently widely used and technologically mature methods. Through the Apriori method, we can delve deeper into the physical and mental health status of men and women and discover their interactions with corresponding evaluation indicators. However, the operational efficiency of this method is often limited by various factors, such as frequent candidate sets and the number of times the database is traversed [4]. To this end, Li T proposed a new Apriori method that can effectively reduce the frequency of candidate sets, better discover their interactions with corresponding evaluation indicators, and more effectively achieve accurate prediction of evaluation results [5]. By adopting hash and object compression methods, the number of candidate options can be adjusted as needed, and important subtasks that have not been considered can be eliminated, effectively improving the method's performance [6]. Through research on the Qin Y algorithm, we found that the Apriori algorithm can be well applied to university laboratory test data. However, one of its drawbacks is that as the scale of laboratory testing expands, the algorithm's performance will gradually decrease [7]. To address this issue, we propose a new algorithm that utilizes transaction compression and hash techniques to optimize the experimental process without the need to pre-determine the size of the samples involved in the experiment [8]. By adopting transaction compression technology, the number and time of traversing the dataset can be significantly reduced, achieving optimal performance of the two algorithms and improving the system's performance. At the same time, it can more accurately obtain the interrelationships between physical measurement data, providing a reliable basis for teachers' educational work to comprehensively and accurately evaluate students' physical and mental health status [9].

2. Apriori Association Rule Algorithm Based on Transaction Compression and Hash Technology

The standards for measuring association rules with support and confidence are as follows. For example, for an association rule like (A'B), the probability of the existence of attribute item A and attribute item B simultaneously is the support of this data association rule, Support (A B). Under the condition of the existence of attribute A, the probability of the existence of attribute B is the confidence of the data association rule, Confidence (A B), which are respectively represented by formula (1) and formula (2).

$$\text{Support}\ (A \Rightarrow B) = P(A \cup B) \qquad (1)$$

Support Count (AUB) represents the number of events that co-occur between A and B, while Total Count refers to the total number of all events.

$$\text{Confidence}\ (A \Rightarrow B) \qquad (2)$$

When A's support program meets or exceeds the minimum requirement of B, that is, A=B, A's reliability will reach an extremely high level. Therefore, support_ Count (A) can be seen as an effective algorithm for strong correlation.

2.1 Apriori Algorithm

The Apriori computation is a method that uses k-itemset iteration. As shown in Figure 1, first, we set a minimum support degree (*min_sup*) and a minimum confidence (*min_Conf*). Then, we scan the entire database, calculate the count of items of total length 1, and filter out items that meet the minimum support degree to obtain the frequent 1-itemset L. Next, we form a candidate 2-itemset (c) based on the links within L, and carry out a pruning process, thereby obtaining the count of items of total length 2 for effective data analysis. By removing items that do not meet the minimum requirements, we get the frequent 2-itemset L2, and continue the iteration until we cannot find the frequent k-itemset *oL*. Finally, we can obtain the frequent itemset L={L1,L2...l}. A drawback of the Apriori algorithm is that its performance declines significantly when the data volume reaches a certain level, which can be seen from the following three aspects: 1) During establishing a connection between L-1 and itself, many candidate items ets may be generated. For example, if *Lk*-1 contains *m* k-itemsets, the time complexity required to establish self-connections will reach O(km^2). 2) When pruning *Ck*, we can determine the fastest scan times by checking whether the *(k-1*)-item subsets of the C candidates are subsets of L1, and the slowest need *k*-1 times, so we can set the average time as *Ck*×*L-1l*×*k*/2. 3) When determining the candidate sets with the lowest confidence, we need to count each item in *C* and scan the entire database.

2.2 Apriori Algorithm Based on Hashing Technique

The Hash technique, also known as the Apriori algorithm, is a hash-based algorithm. Its advantage is that it can compress the generated C, greatly reducing the space occupied by Ck. Firstly, we obtain C by scanning each transaction in the database system and then filter out the candidate item sets that meet the minimum support degree to obtain L. Then, using the hash technique, we can modify the way to form frequent 2-itemsets. Specific contents include forming all 2-itemsets of each transaction and using a hash function to place them in the corresponding hash table for better analysis and processing. Similar information is distributed to multiple buckets, each with its computer. These buckets have their own storage space, significantly reducing storage space consumption. Additionally, by taking similar steps, we can select k-items from the hash table of the k-itemset whose capacity is larger than the support limit, effectively compressing the k-item. The advantage of the Apriori method is that the dataset's characteristics do not limit it, so it can effectively make up for the shortcomings of the hash technology, greatly improving the algorithm's efficiency. In contrast, the disadvantage of hash technology is that it is affected by the average length of transactions. Even if the dataset has fewer characteristics, the average length will correspondingly shorten, resulting in lower operational efficiency. The specific process is shown in Figure 1.

2.3 Apriori Algorithm Based on Transaction Compression Technique

One of the most significant characteristics of the Apriori algorithm is its use of transaction compression methods, which can effectively reduce data processing and improve its capability to handle complex problems. This method's refinement stems from an in-depth study of the traditional Apriori algorithm, considering the unique features of L (*j*>*k*) when processing it for better handling complex issues. By adding and removing specific labels, we can complete the processing of related transactions before scanning the database, greatly reducing the processing frequency. Through the grey relational degree analysis algorithm, we can determine which factors will significantly impact the final result. We will use the Apriori algorithm with the input as *X*={*x*} to better explore these association rules. In the case of .ma, we can use *min_sup* as the lowest support degree threshold. In this way, we can calculate the frequency of *L* in *X*. Specific operational procedures include extracting the necessary information from *x* and C from *X*, as well as corresponding support information; then, we can calculate the frequency of L using the support information of the various subsets of *c* and the result of *min_sup*. Finally, we can calculate the frequency

of 2_ by symmetrical operations on *l*, and calculate the frequency of L by comparing it with *min_sup*. After grey relational degree analysis, the frequency of L significantly decreases, thereby significantly improving its similarity.

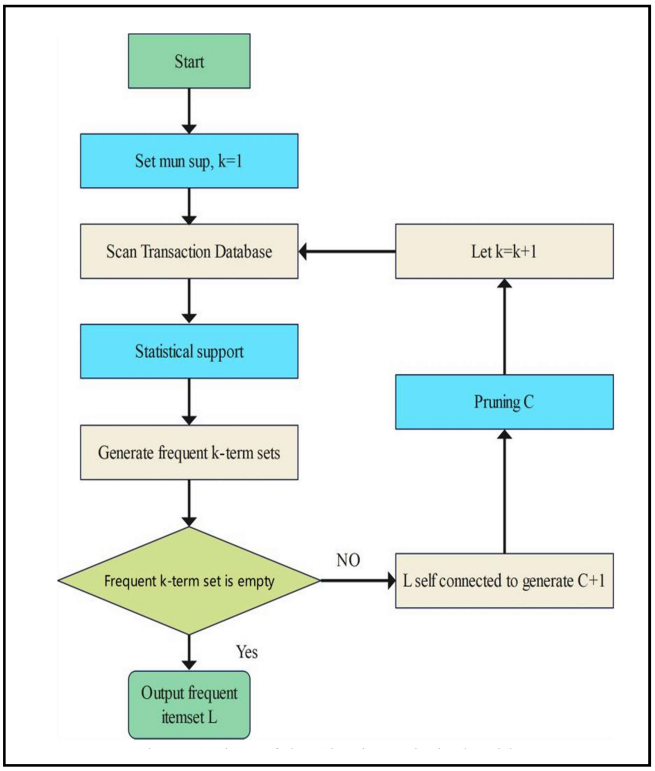

Figure 1. Flow of the Classic Apriori Algorithm

2.4 Improvement Combined with Hashing and Transaction Compression Techniques

Research has found that optimizing Apriori computations relies on the application of hashing techniques, which are beneficial in reducing the size of candidate sets, lowering the frequency of database system retrievals, and improving computational performance. Furthermore, transaction compression technology helps to reduce the size of transactions within the database system, thereby enhancing computational performance. The collaboration between them is also beneficial as their combination can improve the algorithm's performance and avoid any adverse outcomes. Analyzing college physical fitness data, we found that these data have fewer features, but their lengths are the same. We use hashing techniques to traverse the entire transaction set and extract three common itemsets: Li, L2, and L3. Then, we adopt transaction compression techniques to extract L4 from L3 and continue this process until we can no longer find more common k-itemsets. The use of hash techniques can significantly improve its shortcomings.

Using the grey relational degree analysis algorithm, we can transform the original matrix X={*x, xo, ..., x.*} into an average comparison sequence for better analysis and comparison of their relationships... Based on the target sequence *x*, we can calculate the values at each time point of the four comparison sequences *x*, (*t)*}, } = {*xot*, *xo2*, ... ", *xom*} and the comparison of their absolute values to get the minimum value *A*(min) and the maximum value *A*(*max*) of each comparison sequence, thereby determining the minimum value a and the maximum value a of each comparison sequence.

$$\varsigma \text{ ok} = \frac{\Delta(\min) + p\ \ \Delta(\max)}{\Delta_{ok}(t) + P\,\Delta\,(\max)}$$

We can find the maximum value max and the minimum min by calculating the influence ratio between the kth comparison sequence and its reference sequence. If we set the distance between these two sequences as n, we can calculate *nxm*-related ratios, which will be combined to form a matrix. Finally, we can determine their interrelationships by calculating the average of these sequences. After the correlation analysis of the sequences, we found that *Xmax* is one of the important indicators representing the correlation number of the sequence in the *t*period. The size of its value depends on the size of the sequence. Therefore, we take *xmax* as the value of *xmax* to determine the value of max and the size of min, and then determine the size of *m*, so as to determine the size of *nxm* and form the corresponding matrix.

3. Experiment Process and Analysis

3.1 Data Preparation

Through studying the physical examination results of the first semester of 7709 college students, we collected 11 relevant pieces of information, including school code, age, height, weight, vital capacity, 50m run, standing long jump, 1000m run, 800m run, pull-ups, and sit-ups, etc., among which, the information we used comes from the names of these exams. As the data for this experiment come from our physical education courses, their quality is quite reliable. However, some students' PE forms are missing content, and some people may not need to participate. The common steps in data cleaning include: (1) if a student's physical fitness test results are missing items, we should delete the student's results. However, if there are fewer missing items, we can compensate by calculating the average score, mode, or median of the project; (2) for students who did not take the college entrance examination, we can treat any of them as missing items, thus ensuring that the cleaned data is standard, clean, and continuous. Through data information exchange, we can transform continuous data analysis with differences into discrete data analysis with common features, which is more conducive to unified processing. Although the evaluation criteria of each test differ, they are still classified into A, B, C, and D four levels, which helps to carry out tests more orderly and also helps to carry out data mining more effectively. In addition, we also use 1 and 2 to represent male testers and use lowercase letters to define the name of each test; for example, *fhl* is used to represent vital capacity.

By data reduction, we can remove information that may interfere with the algorithm's results while maintaining the integrity of the database. As there are differences in the physical examination content and evaluation criteria for boys and girls, we must handle the data accordingly. Through this approach, we can eliminate information that may disturb the algorithm's results, such as school code and gender.

4. Experimental Verification and Analysis

This experiment uses an Intel(R) Core(TM) i5-5200U 2.20GHz processor, 4GB of memory, Windows 10 operating system, and Python 3.8 programming language. By adjusting the *min_conf* to 0.9 and modifying the minimum support level, we found that even with the Apriori algorithm, we can obtain consistent association patterns with the original algorithm. This indicates that our algorithm meets the requirements and effectively achieves the expected mining accuracy. When the minimum support level is raised to a new level, the

computation speed of our algorithm also significantly improves. By adjusting the reference values, minimum confidence, and minimum support, we observed performance improvements in our algorithm and found it more cost-effective than traditional algorithms. Compared with other algorithms, we found that our algorithm performs similarly to traditional ones and exhibits good stability. Apriori algorithm performs well in handling complex data but has slightly lower efficiency. On the other hand, the hash algorithm performs well in data processing, providing faster processing speed and more accurate prediction capabilities. Therefore, in this paper, we propose the hash algorithm and demonstrate its ability to enhance processing capabilities through its application.

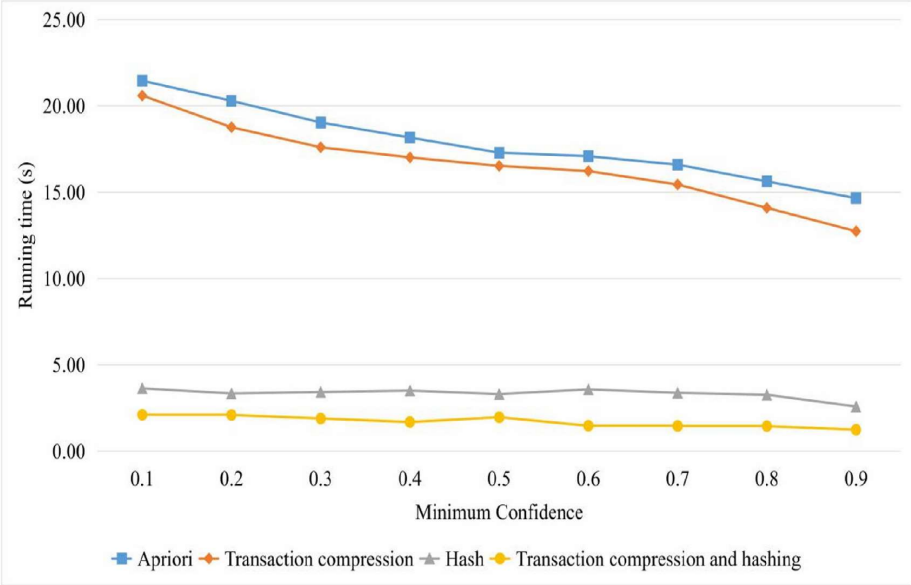

Figure 2. Performance Comparison with Different Confidence Levels

After preprocessing the original physical examination data, we obtained 2,266 records of female physical examination data and 5,071 records of male physical examination data. Assuming *min_conf* = 0.9 and *min_sup* = 0.3, we used four different methods to perform association rule data mining on the two types of information and compared their performance (Figure 2). Our proposed method, which combines transaction compression and hash, outperforms other methods, showing excellent accuracy and stability and achieving good performance. Through gender analysis, we discovered that when using the Apriori method for physical examination, the efficiency of females is significantly higher than that of others. The efficiency for females improved by 86.12%, while for males, it increased by 90.63%. Furthermore, when using the transaction compression algorithm, the efficiency for females is also significantly higher than that for males, with an improvement of 80.57% and 89.32%, respectively. In the case of using the hash algorithm, the efficiency improvement for females is 48.1%, while for males, it is 55.69%. After applying the four different algorithms, the experimental results for males showed a decisive advantage, which becomes more pronounced as the data analysis set increases .

Based on Figure 3, we obtained valid data for analyzing the mutual influence between male and female physical examination scores. We found that the first relevant rule, "*50m:C'ytxs:D*" has a support of 72.1% and a confidence of 93.5%. Therefore, the data indicates that when a student's 50m exam score is at the C level, it is very

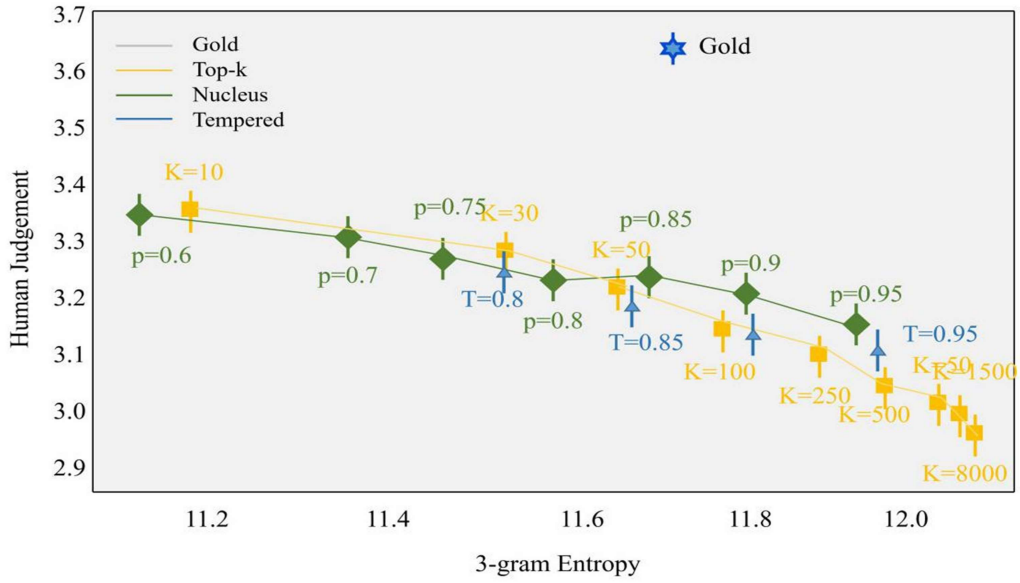

Figure 3. Comparison of Efficiency for Four Algorithms

likely that their pull-up score will reach the D level, with a probability of 72.1%. Using this example, we can utilize a student's historical records to estimate their scores in various exams, providing timely reminders and encouraging their development in specific areas. This can assist teachers in formulating more practical and valuable training plans.

5. Conclusions

With the advancement of technology, data mining techniques have become increasingly sophisticated, but their application in sports is still limited. In this study, we used four different methods, including third-party computation and Apriori method, to evaluate their advantages and disadvantages in physical examinations for college students. When dealing with large-scale data, we found that our method excels in execution efficiency, while the winner's advantage lies in accuracy, robustness, and robustness. Through in-depth research, we discovered the mutual influence between physical examination results and other factors, revealing students' health conditions and providing reliable references for their healthy development, thus offering valuable suggestions.

References

[1] Guo, H., Liu, H., Chen, J. Y., et al. (2021). Data Mining and Risk Prediction Based on Apriori Improved Algorithm for Lung Cancer. *Journal of Signal Processing Systems*, 1-15.

[2] Peng, S., Xie, X., Zhai, J., et al. (2021). A Page-topic Relevance Algorithm Based on BM25 and Paragraph-Semantic Correlation. *Journal of Physics: Conference Series, 1757*(1), 012115 (4pp).

[3] Widyastari, D. A., Kesaro, S., Rasri, N., et al. (2022). Learning Methods During School Closure and Its

Correlation With Anxiety and Health Behavior of Thai Students. *Frontiers in Pediatrics, 10*, 815148.

[4] Pizzulli, V. A., Telesca, V., Covatariu, G. (2021). Analysis of Correlation between Climate Change and Human Health Based on a Machine Learning Approach. *Healthcare*, (1).

[5] Xie, H. (2021). Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets. *Open Journal of Social Sciences, 09*(4), 458-468.

[6] Lian, Z., Liu, S., Lu, J., et al. (2021). Research on Sports and Health Intelligent Diagnosis Based on Cluster Analysis. *Scientific Programming*, (13), 2021.

[7] Li, T. (2021). Application of APRIORI correlation algorithm on music education curriculum association rules. *Journal of Physics: Conference Series, 1955*(1), 012067 (6pp).

[8] Qin, Y., Wang, J., Qin, J., et al. (2022). Correlation Meta-Analysis of the Efficacy of Inhaled Corticosteroids in Children with Asthma Based on Smart Medical Health. *Journal of Healthcare Engineering, 2022*, 6220774.

[9] Fan, C. (2021). The Physical Health Evaluation of Adolescent Students Based on Big Data. *Mobile Information Systems, 2021*, 1-6.

[10] Pan, T. (2021). An Improved Apriori Algorithm for Association Mining Between Physical Fitness Indices of College Students. *International Journal of Emerging Technologies in Learning (iJET), 2021*(09), 22747.