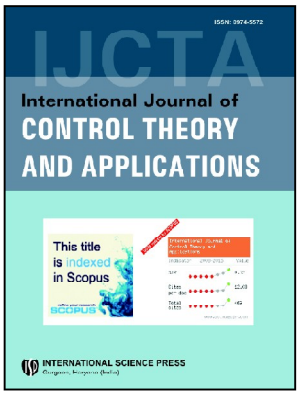

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 11 • 2017

Efficient Privacy Preservation High Utility Itemset Mining for Large Data Responses

Amandeep Kaur and Manjot Kaur

Chandigarh University, E-mail: akaur6015@gmail.com, manjot.cse@cumail.in

Abstract: The data retrieval processes always pose the threat of the illegal and unauthorized data extraction from the online databases through the hacking tricks applied through the user queries. These data breaches are helpful for the hacking elements for the purpose of data stealing from the server. This sensitive information contains the sensitive factors such as financial details of the personals, the companies and other profitable or security related organizations. The aim of this research is to build the automatic content data filtering engine, which works on the concept of the high utility mining. The high utility mining techniques are associated with the recognition of the sensitivity of the itemsets higher than the input query. The high utility itemsets may contain the information of higher level than the scope of the user's query, which is why it has to be protected. The major idea to protect the data over the online datasets that stays higher than the threshold limit computed using the high utility mining privacy protection algorithm. The proposed model aims at the evaluation of the high utility mining itemset in order to filter the highly sensitive content in the response to the user's input query. The proposed model has been designed to prevent the data leakage caused by the attacks by using the high utility queries which are mostly not permitted from the user's end. The proposed model has been analyzed for its capability to protect the sensitive information from the low utility mining queries. The experimental results have proved the efficiency of the proposed model in terms of data classification, hiding failures and many others.

Keywords: High Utility Itemsets (HUI), Privacy preserving mining, Association rule based filtering, Information sanitization.

1. INTRODUCTION

Data mining outlined as finding hidden info from giant knowledge sources has become a well-liked thanks to discover strategic information. unsolicited mail selling, internet site personalization, bioinformatics, mastercard fraud detection and market basket analysis square measure some examples wherever data processing techniques square measure normally used. data processing tools predict future trends and behaviors, permitting businesses to form proactive, knowledge-driven choices. data processing finds valuable info hidden in giant volumes of information. It's the analysis of information and also the use of software package techniques for locating patterns and regularities in sets of information. we are able to simply notice the hidden patterns from great amount knowledge mistreatment data-mining. the goal of {the data|the info|the info} mining method is to extract information from an information set and remodel it into an obvious structure for additional use.

Association rule mining is one in every of the foremost vital and well researched techniques of information mining. for instance take into account a grocery with the massive assortment of things. Management call of grocery includes things available, planning of coupons, most profit with reference to merchandised. Previous transactions used for enhancements of call. Bar codes helps to store the things purchased on per-transaction basis known as as basket. Mining association rules from an outsized information, has been a crucial task within the space of information mining to get hidden, attention-grabbing associations that occur between varied knowledge things. today ARM is broadly speaking utilized in many alternative areas like telecommunication networks, market and risk management, internal control mobile mining, graph mining, academic mining, etc.

Association Rule concealing is that the method of concealing sturdy association rules and making sanitised information from the first information so as to stop from unauthorized access. Association rule concealing refers to the method of modifying the first information in such how that sure sensitive association rules disappear while not seriously touching the information and also the non-sensitive rules.

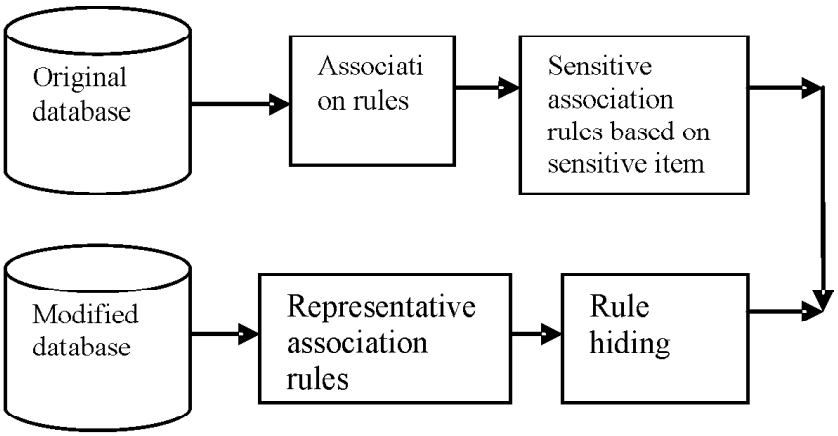

Figure 1: Framework of Association Rule Hiding [6]

The association rule concealing drawback are often thought-about as a variation of the well-known information reasoning management drawback in applied mathematics and construction databases. the first goal, within the information reasoning management, isto defend access to sensitive info that may be obtained through non-sensitive knowledge and reasoning rules. The main purpose of the association rule concealing algorithms is to form the sensitive rules invisible which may be generated by association rule mining algorithms.

1.1. Privacy Preservation data processing

In recent years, advances in hardware technology and also the growth of net have semiconductor diode to a rise within the capability to store and record personal knowledge concerning customers and people. This has semiconductor diode to considerations that the private knowledge is also put-upon for a range of functions. so as to alleviate these considerations, variety of techniques have recently been planned so as to perform the information mining tasks during a privacy-preserving approach. The techniques used for acting privacy-preserving data processing square measure drawn from a large array of connected topics like data processing, cryptography and knowledge concealing. whereas an outsized variety of analysis papers square measure currently accessible during this field, several of the topics are studied by completely different communities like information community, the applied mathematics revealing community and also the cryptography community with completely different designs. Privacy conserving data processing has become very hip for safeguarding the confidential information that has been extracted from the information mining techniques.

1.2. Objective

The main objective of privacy conserving data processing is to develop algorithms for modifying the first knowledge in a way. A primary necessity of privacy-preserving data processing is to guard the input file, however enable knowledge miners to extract helpful information models. Privacy is one in every of the foremost essential properties that associate degree data system should satisfy. For this reason, varied efforts are created to integrate privacy conserving techniques with data processing algorithms so as to avoid the revealing of sensitive info throughout the information discovery.

The planned formula is that the improved version of MDSRRC. MDSRRC couldn't hide association rules with multiple things in L.H.S and multiple things in R.H.S. therefore the rule is like aX→bY wherever a, b→ I and X,Y→I. Here b is associate degree item chosen by planned formula to decrease the support of the R.H.S. and reduce the boldness of the rule below MCT. to beat this limitation, we've got planned a sensitive rule filtering (SRF) FP-Growth formula that has been designed to unravel the matter of planning the correct info concealing rules.It modifies the minimum variety of transactions to cover most sensitive rules and maintain knowledge quality.

2. RELATED WORK

Domadiya *et al.* (2013) planned that a heuristic primarily based rule named MDSRRC (Modified Decrease Support of R.H.S. item of Rule Clusters) to cover the sensitive association rules with multiple things for the utilization of (R.H.S) and antecedent (L.H.S). This rule overcomes the limitation of existing rule concealment rule DSRRC. The planned rules are used for selection of the things and data transactions supported bound criteria that modify transactions to cover the sensitive info.

Koh and Shieh (2004) planned a general progressive change technique for maintaining the frequent itemsets discovered in an exceedingly info within the cases together with insertion, deletion, and modification of transactions within the info. associate degree economical rule, known as AFPIM (Adjusting FP tree for progressive Mining), is intended supported adjusting FP-tree structures. This approach uses a FP-tree structure to store the compact info of transactions involving frequent and pre-frequent things within the original info. In most cases, with no need store scan the initial info, the new FP-tree structure of the updated info is obtained by adjusting FP-tree of the initial info per the modified transactions.

Bhat *et al.* (2014) proposing the heuristic primarily based rule for concealment the sensitive association rules the rule is known as as MDSRRC , owner hide sensitive association rule and place remodel rules to the server for outsourcing purpose. during this rule they supply associate degree progressive association rule for mining. The Matrix Apriori rule is planned that relies on analysis of 2 association rule named as Apriori rule and FP-growth rule. The matrix Apriori rule contains a easy structure similar as a matrices and vectors; the rule generates frequent patterns and minimizes the quantity of sets,as compared to previous rule. The matrix rule is easy and economical thanks to generate association rule than the previous rule. For concealment the sensitive info of the info planned rule MDSSRC selects the transactions and things by victimisation bound criteria that remodel.

Lan *et al.* (2009) provides that frequent itemsets mining plays a vital role in association rules mining. The apriori rule and therefore the FP-growth rule ar the foremost noted algorithms, existing frequent itemsets mining algorithms ar virtually improved supported the 2 algorithms severally and suffer from several issues once mining huge transactional datasets. a replacement rule named APFT is planned, it combines the Apriori rule and FP-tree structure that planned in FP-growth rule. The Advantage of APFT is that it does not have to be compelled to generate conditional pattern bases and sub-conditional pattern tree recursively and therefore the results of the experiments show that it works quicker than Apriori and virtually as fast as FP-growth.

Agrawal *et al.* (1994) presents 2 new algorithms for locating the association rules between things of enormous datasets of sales transactions known as Apriori and Advanced Apriori. Empirical analysis shows that these algorithms outmatch the famed algorithms by factors starting from 3 for tiny issues to quite associate degree order of magnitude for big issues. Integrated rule known as ad Apriori Hybrid. Apriori Hybrid has outstanding properties with regard to the dealing size and variety of things in info. The execution time decreases somewhat because the variety of things within the info will increase.

3. PERFORMANCE MEASURES

There are several parameters to measure the performance of the database sanitization algorithms. The proposed model performance has been evaluated using the following performance parameters:

Hiding Failure: The parameter of the hiding failures measures the effectiveness of the restrictive rule definitions by analyzing their appearance in the response database after the application of sanitization. The hiding failures are computed by evaluating the sanitized data against the sensitive rule population.

$$HF = \frac{|SR(D')|}{|SR(D)|} \tag{3.1}$$

Artifactual Patterns: The count of the legitimate sanitization rule defined upon the basis of the association rule mining in order to filter the noise from the data. The rate of the mined artifact from the given data is computed as the artifactual patterns, which is measures by using the following equation on the basis of mined association rules (given by P) over the training dataset (D).

$$AF = \frac{|P^t| - |P \cap P^t|}{P^t} \tag{3.3}$$

The algorithm evaluates the D against the sanitized dataset D' for the computation of effectiveness of the P'.

Misses Cost: The percentage of the high utility itemsets, which accidently appears in the final product of the results, is known as the misses cost. The misses cost parameter evaluates the overall performance of the proposed model in the terms of computational errors for the sanitization of the high utility itemset mining (HUIM).

$$MC = \frac{|SR(D')| - |SR'(D')|}{|SR^t(D)|} \tag{3.2}$$

Difference: The evaluation of the percentage of the difference between the sanitized dataset and the original input dataset. The difference is computed by the following equation:

$$Diff(D, D') = \frac{1}{\Sigma_{i=0}^{n} f\,d(i)} \times \Sigma_{i=1}^{n} [f\,d(i) - f\,d(i)] \tag{3.4}$$

Where the frequency of the itemset (i) is represented with the fd(i) from the original database, where the itemset i of the sanitized database is given by fd'(i). n represents the number of total sanitization filter.

4. PROPOSED MODEL

In the proposed model, we have aimed at the development of the effective data sanitization algorithm for the filtering of the high utility itemsets from the user's query over the cloud platforms. There are several kinds of the sensitive data saved on the cloud platforms such as credit card information, economic data, company accounts, personal accounts, etc which must be protected in order protect the financial and personal information of its

users and organizations in order to realize the high-order portal security. The proposed model has been designed specifically to filtering the high utility itemsets (HUI) by analyzing the level of sensitivity of the queried itemsets against the input user query.

The existing algorithm has been improved by using the balanced and dynamic mining rule generation algorithm. The dynamic information optimization for the higher level of information combination retrieval to achieve the higher accuracy in the complex data. The ideal multi-thread analysis over the Database Structure Similarity (DSS), Database Utility Similarity (DUS) and Itemset Utility Similarity (IUT) must be performed with in-depth analytical engine in order to improve the overall accuracy of the existing model. The DSS provides the initial stage similarity between the user query and the itemsets. The proposed model incorporates the calculation of the higher order similarity in the next step by utilizing the database utility similarity (DUS) and itemset utility similarity (IUS) after evaluating the high utility itemests out of the matching data itemsets. The collaborative indexes model can improve the overall quality of the sensitive information sanitization. The collaborative index will be created by combining the DSS, DUS and IUT indexes, which gives the normalized sensitivity in order to retain the maximum non-sensitive information. The collaborative approach helps us to recreate the SHUI (sensitive/secure high utility itemests) list, which contains the vital information for the sanitization purposes. The rule-based mining approach has been utilized to minimize the sanitization errors. The fidelity of the information is preserved by using the effective sanitization algorithm, which primarily utilizes the definitive module to create the filtering rules by analyzing the sensitivity of the itemsets by using the High Utility mining (HUM) methods. The various indices are computed altogether to determine the sensitivity of the given itemset against the query input.

Algorithm 1: Hybrid High Utility Filtering Algorithm (HHUFA)

INPUT

- Input dataset (Id)
- Rule Sensitivity (Rs)
- Filtering rule data (Frd)

OUTPUT

- Sanitized data (Sd)

MAIN ALGORITHM

1. Acquire the input dataset (Id) in the runtime memory
2. Acquire the input query sample
3. Load the initial cost dataset
4. Compute the frequency of the individual itemsets (itemset utility) and cumulative row utility
5. Shortlist the high frequency itemests before the previous
6. Evaluate the individual itemset frequency in the form of the high utility itemests
7. Compute the database structure similarity (DSS)
8. Elaborate the **Itemset Utility Similarity (IUT) index over the DSS**
   a. Return the shortlisted dataset using IUT
9. **Elaborate the itemset utility using the database utility similarity (DUS)**
10. Estimate the secure high utility itemsets using the SHUI model.
11. Generate the itemset mining threshold known as T

12. Shortlist the high utility itemsets under the SHUI model
13. Filter the high utility itemsets (HUI) from the database query using the high utility itemset mining (HUIM) model.
14. Create the information sanitization according to the sanitization rule (SRu)
15. Prepare the SHUI for each itemest and each row
16. Run the iteration on all of the database rows
17. Run the iteration for all columns in the current row
    a. If the entity has higher utility than the rule SRu
        i. Filter the itemset
    b. If its not the last column
        i. If its not the last row
        1. GOTO 17
    c. Otherwise
        i. Return the sanitized dataset.

5. RESULT ANALYSIS

The visualization of the dataset results has been incorporated using the high utility dataset mining for the privacy preserving among the large databases hosted over the cloud computing resources. The visualization of the data includes the two primary groups defined with the yellow colored disks and the black colored marker with plus sign. The black colored plus sign markers define the itemsets with low utility mining (also called low utility mining itemsets) and the yellow colored markers define the high utility itemsets, which are placed in the higher scope according to the queried arguments. The high utility itemsets must be recognized before the packaging the

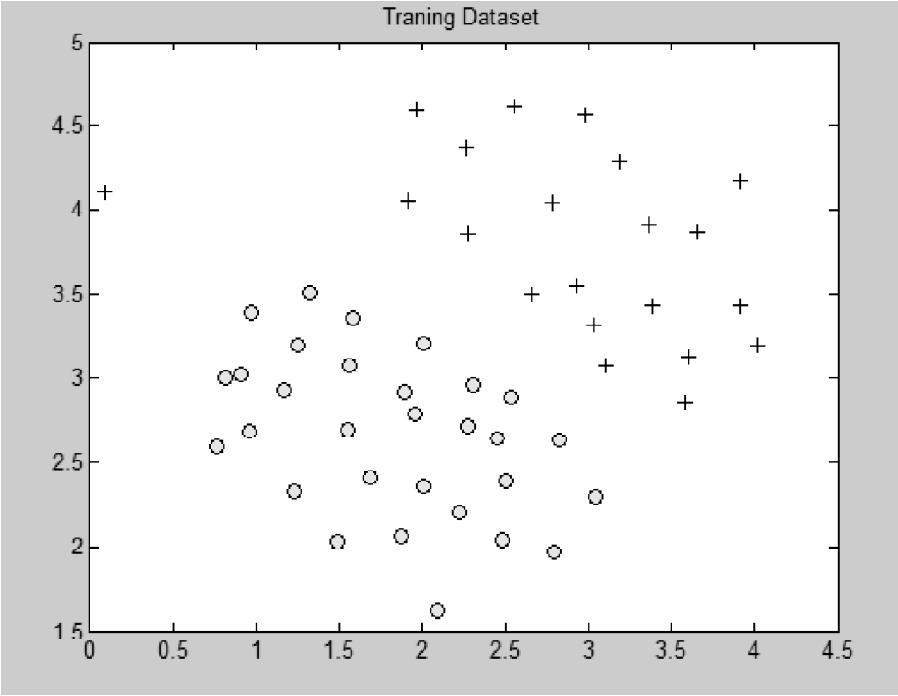

Figure 5.1: Dataset 1 before applying HUI mining

presentation of the final output result over the database. An ideal privacy preserving model must be capable of recognizing the itemsets with the under utility and higher utility in order to produce the appropriate results to the end users. This section visually shows the results obtained from the proposed model.

The figure 5.1 describes the high utility itemsets and the under utility itemsets. The proposed model aims at the classification of the itemsets according to the input arguments. The proposed model has been applied over the input data to recognize the scope of the given itemset data.

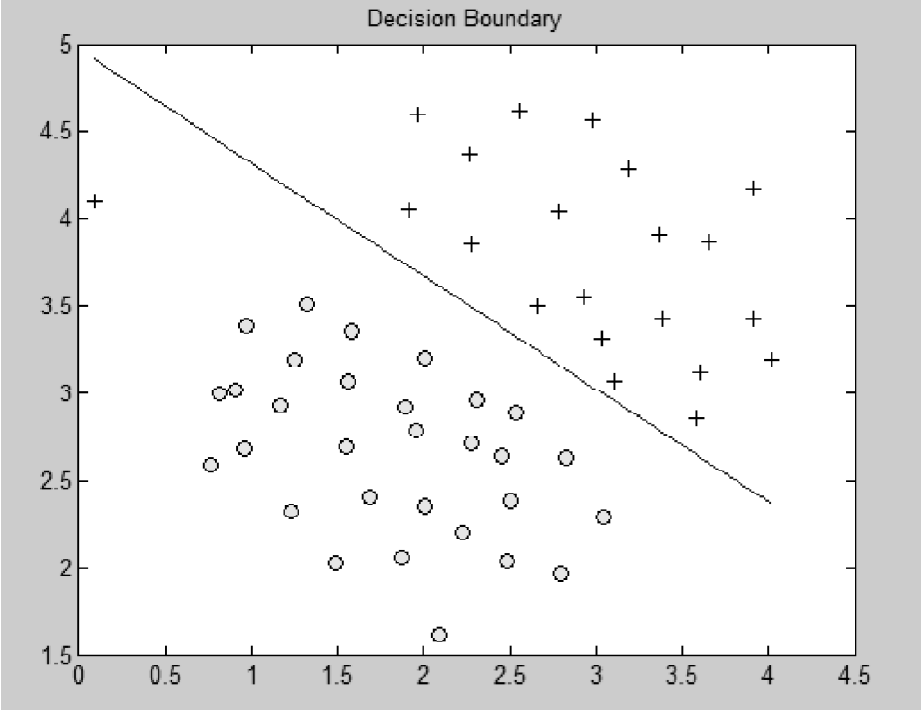

Figure 5.2: After the incorporation of the decision boundary using the HUI mining on dataset 1

The figure 5.2 describes the results after the incorporation of the privacy preserving high utility itemset mining. The proposed model results show the clear picture in terms of recognizing the high utility itemsets in the given dataset. The distant sub-datasets have been combined in this scenario, where the under utility limit and high utility itemsets are significantly described as shown in the above figure 5.2. The algorithm decision has been represented by calculating the decision boundary, which is represented in the blue line. The algorithm decision has calculated properly and shows the proper significance of the proposed model in recognizing the high utility itemsets with the proposed HUI mining scheme. In the another scenario (Scenario 2), the distorted itemsets have been collected and grouped altogether. The minimum distance has been kept between some of the itemsets in both of the groups including the high utility and under utility limits as per shown in the figure 5.3. The proposed model application has been applied over the given dataset and the results are recognized from the output results in the figure 5.4.

The figure 5.3 describes the high utility itemsets and the under utility itemsets with low distantia factoring. The low distantia or distance factor includes the border line itemests which are mostly the part of the itemesets grouped under the misses costs or hiding failures. The proposed model aims at the classification of the itemsets according to the input arguments. The proposed model has been applied over the input data to recognize the scope of the given itemset data. The crucial part of the this scenario includes the incorporation of the privacy preservation based HUI mining along with the input argument based decision making process.

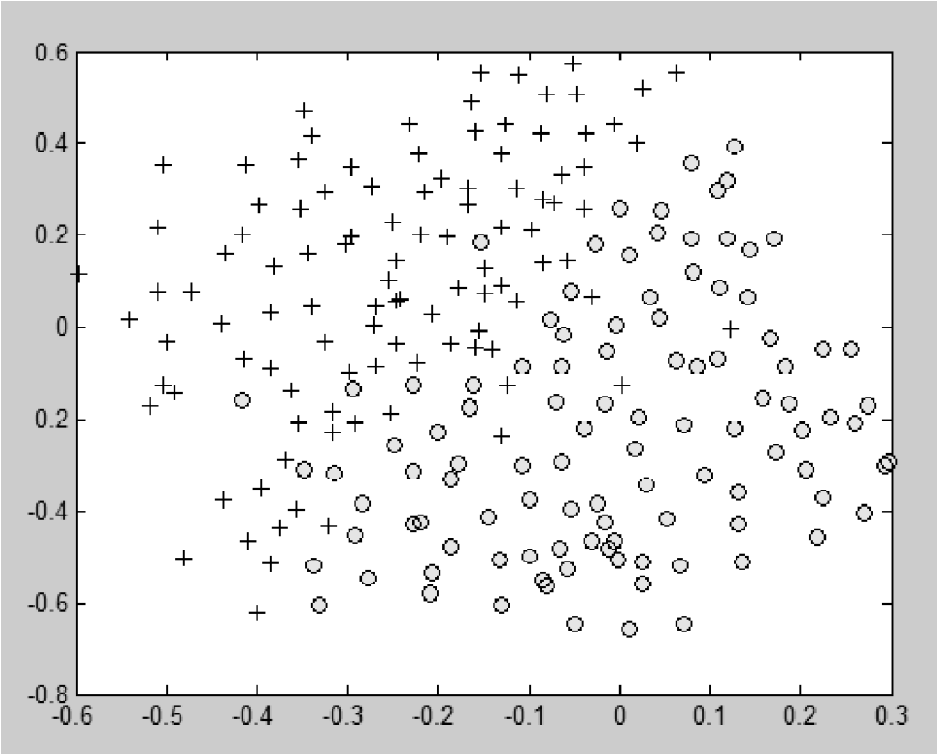

Figure 5.3: Dataset 2 before the application of the HUI mining

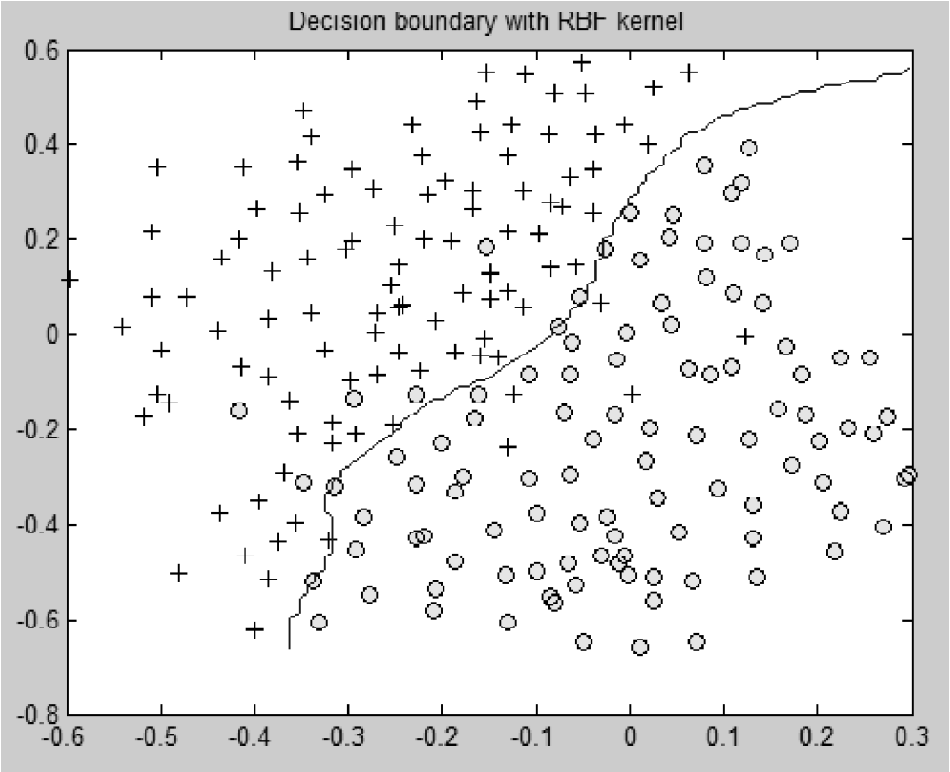

Figure 5.4: Dataset 2 after the incorporation of the decision boundary calculation using the HUI mining

The figure 5.4 describes the results after the incorporation of the privacy preserving high utility itemset mining. The proposed model results show the clear picture in terms of recognizing the high utility itemsets in the given dataset.

The table 5.1 contains the performance measures for the evaluation of the proposed model information sanitization method. The proposed model has been evaluated in the form of the primary aspects of the itemset filtering measures.

Table 5.1
The evaluation of the proposed information sanitization algorithm

Artifactual Patterns	0%
Difference	3.25%
Hiding Failures	0%
Misses Cost	16.12%

The performance factors from the table 5.1 have been evaluated using the proposed model, which are compared against the existing model in the following table 5.2. The performance measures have been evaluated in the table 5.2 against the results obtained from the existing model.

Table 5.2
Overall performance measurement of the proposed model in comparison with the existing model

Error Type	*Proposed*	*MDSRRC*	*SRFFP*
Hiding Failures	0%	0%	0%
Misses Cost	20%	26.66%	20%
Artifactual Patterns	0%	0%	0%
Difference	4.17%	5.4%	4.17%

6. CONCLUSION

The distant sub-datasets have been combined in this scenario, where the under utility limit and high utility itemsets are described with minimum distance as shown in the above figure 5.4. The algorithm decision has been represented by calculating the decision boundary, which is represented in the blue line. The algorithm decision has calculated and shows that the most of the itemsets classified properly with higher significance, whereas some of the itemsets has been classified as the decision misses cost or hiding failures while applying the proposed model in recognizing the high utility itemsets with the proposed HUI mining scheme. The proposed model has been evaluated for its performance to determine the high utility itemsets in the given user query. The experimental results have proved the efficiency of the proposed model in determining the high frequency itemsets with the precision. However, there is still the requirement of the critical improvement to handle the nearest entity separation, when they belong to the distinct categories. Hence, the proposed model can be improved by using the micro-and-macro threshold based factoring of the itemest data in order to filter the high utility itemsets more precisely.

References

[1] Agrawal R, Srikant R., "Privacy-preserving data mining" Proceedings of the ACM SIGMOD Inter-national Conference on Management of Data; Dallas, Texas; 2000, 439–450.

[2] Atallah M., Elmagarmid A., Ibrahim M. Bertino E., and Verykios V., "Disclosure limitation of sensitive rules," in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX'99. Washington, DC, USA: IEEE Computer Society, pp. 45–52, 1999.

[3] Bertino E, Lin D, Jiang W., "A survey of quantiûcation of privacy preserving data mining algorithms, In "Privacy-Preserving Data Mining—Models and Algorithms. Advances in Database System. Springer, Berlin,34:183–205, 2005.

[4] Bhat P., Malviya M., and Lade S., "Optimization of MDSRRC with Matrix Apriori", International Journal of Operations and Logistics Management, Vol. 3, Issue: 2, pp 140-147, 2014.

[5] Cheung D., Ng V., Fu A., Fu Y., "Efficient Mining of Association Rules in Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 8(6), pp 911-922, 1996.

[6] Dhutraj N., Sasane S., Kshirsagar V., "Hiding Sensitive Association Rule for Privacy Preservation", Institute of Electrical and Electronics Engineers (IEEE) Transactions on knowledge and data engineering, 2013.

[7] Domadiya N., RaoU., "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database", Institute of Electrical and Electronics Engineers (IEEE), pp. 1306-1310, 2013.

[8] Fukuda T., Morimoto Y., Morishita S., Tokuyama T., "Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization", ACM SIGMOD international conference on Management of data, Volume 25 Issue 2, pp 13-23, 1996.

[9] Ghosh A., Nath B., "Multi-objective rule mining using genetic algorithms", Soft Computing Data Mining Elsevier, Volume 163, Issue 1-3, pp 123-133, 2004.

[10] Kaur C., "Association Rule Mining using Apriori Algorithm: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 2(6), pp-893-900, 2013.

[11] Koh J. and Shieh S. (2004), "An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-Tree Structures",in LNCS 2973, pp. 417–424, 2004.

[12] Lan Q., Zhang .D,Wu B.," A New Algorithm For Frequent Itemsets Mining", Institute of Electrical and Electronics Engineers (IEEE), pp. 360-364.

[13] Tanbeer, Syed Khairuzzaman, Chowdhury Farhan Ahmed, Byeong-Soo Jeong, and Young-Koo Lee. "CP-tree: a tree structure for single-pass frequent pattern mining." In *Advances in Knowledge Discovery and Data Mining*, pp. 1022-1027. Springer Berlin Heidelberg, 2008.

[14] Totad, Shashikumar G., R. B. Geeta, and PVGD Prasad Reddy. "Batch incremental processing for FP-tree construction using FP-Growth algorithm." *Knowledge and information systems* 33, no. 2 (2012): 475-490.