International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

A Novel Filtering Based Scheme for Privacy Preserving Data Mining

Charu Sharma[1], Dr. Kanwal Garg[2]

Research Scholar[1], Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
Assistant Professor[2], Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India

---

Abstract- *Now a day, there exists a pressure in sharing the personal information, and it raises an issue related to the seclusion of data. When the data is extracted from various sources or parties, then a concealment concern arises that forbid the data from directly being shared. So, this paper addresses the filtering based algorithm for mining of noisy and unclean data, and this results in providing the sanitized data that does not contain any sought of redundant values, strings (personal information). The filtering is done in categorical data and textual instead of numerical data. Hence, the target of paper is to implement an innovative filtering based algorithm for seclusion of data that maintains data utility and has no information loss. This paper compares the final results for the time required and some rules mined per Data set.*

Keywords: ***-Data Mining, Feature Subset Selection, Information Loss, K-Anonymization, Privacy, ReliefF, Re-Identification, Security****.*

I. Introduction

Data Mining is a technology that results in finding fruitful patterns or proficiency from a large amount of database. The patterns or knowledge discovered to contain a certain amount of sensitive information about an individual or an organization. Privacy Preserving Data Mining consider the problem of maintaining the confidentiality of data whereas various PPDM techniques are used to alter the original data in a way that no private information is leaked and is protected from attacks [1]. Association Rule Hiding is a privacy preserving method that is used to hide sensitive association rule. The major area of concern does also exist as some non-sensitive data can also deliver confidential information. The primary goal of privacy in data mining is to build algorithms for transforming the original data into secured/unsecured way. The PPDM technique divides into two broad fields:

1. Information concealment
2. knowledge masking

Data suppressing is the elimination or alteration of super sensitive information from the data before disclosing it to others whereas Knowledge masking focuses on concealing the sensitive knowledge which can be excavated from the database using any data mining algorithm. The problem of hiding association rule considers to be a type of database inference control, but its prime intent is to protect the touchy rules, not the raw data [2].

1) Association Rule Mining

It is one of the privacy preserving methods that are used to protect sensitive association rules. It creates a sanitized database from the original database so that the unauthorized party could not generate typical delicate patterns. It provides the user to read or can only access non-sensitive rules [3]. It scans the whole transaction and compute the support and confidence of the rules and recover only those rules whose support and confidence is higher as compared to minimum support and confidence threshold. It is a two-step process:

1. To find all the common items which appear at least as frequent as a pre-determined count of minimum support.
2. Render the rules based on minimum support and confidence [4].

Let us suppose a given transactional database 'D,' having minimum support and minimum confidence and set R to be the mined rules from database 'D.' Let 'RAsen' be the subset of R which denotes a set of sensitive association rules which are supposed to hide. The principal objective is to find the sanitized database in such a way that all 'RAsen' fragments will remain protected, while a set of 'RAnon-sen' will be minimum. As we apply data mining technique on sanitized database, the 'RAnon-sen' gets divided into two controls: Association rule and lost rules whereas 'RAsen' will get split into the group of sensitive rules that are not hidden (R_{nh}) and 'R_h' that is a collection of hidden sensitive rules [5]. If any of the rule having a support > k globally, then it must have a support > k on at least one of the respective sites. The algorithm would work as follows: It makes a request to all the sites to direct all the rules with the support of at least k. For every reoccurrence of rule it directs all the sites to address the count of their transactions that support the rule, and the total number of transactions at the site. So, this criteria

computes the global support of each rule and provides all the rule that have a minimum support of at least k [5, 6].

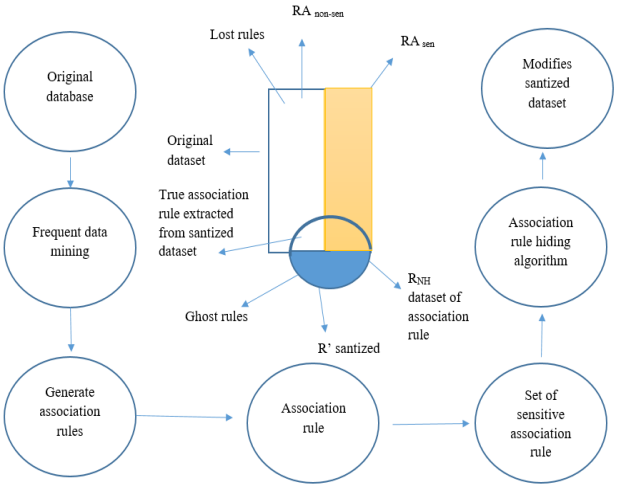


Fig 1. Association rule mining process [5]

2) K-anonymity

K-Anonymity is the privacy preserving data mining technique that helps in delivering a tremendous amount of data so that it can be used for business or research by various organizations by ensuring that no privacy will be leaked related to an individual and would not put the released data in danger. It protects the released data against interpretation and linking of attacks [7]. The two basic definitions related to k-anonymity corresponding to data.

1. Key Attribute - This attribute identifies an individual directly, and this data is removed at the time of release. E.g. Name, Mobile number.
2. Quasi-Identifier - This identifier is used to represent set of attributes that identify an individual is called quasi-identifier. E.g. Date of birth, Pin code.
3. Sensitive Attribute - Those attributes that contain the confidential information about an individual. E.g. Salary, Health problem [7].

The two main ambition for privacy protection are to safeguard the individual identifications and to shield the sensitive relationships. With k-anonymity, the master data set comprising the confidential data can be reconstructed such that it will be troublesome for an attacker to regulate the individuality of an individual. The maximum probability of any record in k-anonymized data set is 1/k [8]. There are two re-identification scenarios for a specific individual:

1. Re-identify a specific individual: An intruder knows that a particular individual belongs to that particular anonymized data set and wishing to find the record that belongs to that particular individual.
2. Re-identify an arbitrary individual: An attacker is not interested in knowing the person who is re-identified. The intruder is only interested in claiming and disclosing the data to the organizations [8].

3) Relieff

The Relieff algorithm is used for frequent subset selection and estimates the weights of the attributes in the dataset. Feature subset selection is an approach that is used for reducing the attribute space in the data set. It identifies a fragment of features by removing redundant data [9]. The valuable feature set contains a remarkably consistent feature which helps in improving the efficiency of the algorithms to separate them precisely. Relieff is a feature selection algorithm that is used for feature weight calculation of random instances.

Feature subset selection is the method that identifies and removes a lot of extraneous and redundant features. Therefore, the inappropriate features do not give the predictive correctness and the non-essential features do not return a superior predictor for the given primary data. For Machine learning techniques the feature selection is applied at the level of data preprocessing [10].

It aims to find the number of functions that describe the dataset in a better way than the original data set. Feature subset selection provides the support to deal with the "curse of dimensionality" problem by eliminating the inappropriate and identical features. It speeds up the learning algorithms and improves the efficiency and performance of the algorithm [10].

The chief diversity between multi-label and single-label learning is that the class values are usually correlated, whereas the class values in single-label are contradictory. Therefore, Multi-label learning is emerging as a research topic due to its increase of use in number of applications example such as bioinformatics, text mining, etc [11].

II. Related Work

Aldeen, Y.A.A.S. et. al [1] this paper provides the basic knowledge of preservation of data by the use of various data mining techniques for the use of data in mining purposes, the quality of data is maintained and confidentiality of the information is preserved.

Patel, D.S. et. al [2] it provides the brief view of the data mining techniques that are used for preservation of data. This paper presents the solution to the above problem by using cryptographic technique to mine the data with privacy.

Jaideep Vaidya, Chris Clifton et. al [3] addressed the issue of association rule mining where the transactions are shared among various sites and identify the valid global association rules. However, the site does not reveal any personal information and presented a two-party algorithm that is used to mine frequent item sets with minimum support with confidentiality.

Murat Kantarcıoglu, Chris Clifton et. al [4] addressed a secure mining of association rules over horizontally partitioned data and these methods use cryptographic techniques for minimizing sharing of information and adding a small amount of overhead to the data mining task. Kharwar Ankit, R. et. al [5] this paper provides an existing approach for association rule hiding and provides a survey on heuristic based algorithms and various techniques to generate useful pattern by hiding sensitive rules so that confidential information is not disclosed to any of the third party.

Verykios, V.S. et. al [6] this provides the exact, novel edge-based approaches for association rule hiding that is used to achieve a useful pattern after mining of data by hiding frequent item sets.

Padmapriya G. et. al [7] evaluated the re-identification risk of anonymization technique and the improvements on three massive data sets. For one of the re-identification scenarios, it performs over-anonymization within the small sampling fragments. Over-anonymization results in enormous misinterpretation in data, which makes the data less profitable.

Khaled El Emam, Fida Kamal Dankar et. al [8] proposed a premise testing approach that provides a perfect control over re-identification risk and cut back the intensity of information loss as compared to baseline k-anonymity.

Monard M. C. et. al [9] proposed a new extended version of single label feature selection algorithm i.e. Multi-label feature selection or ReliefF algorithm. The multi-label feature subset selection strictly performs one dimensional measures for predictor ranking and the consequence of interacting dimensions that deal with multi label data without any modification of data.

Durgabai, R.P.L. [10] in this paper an algorithm is proposed for minimizing the errors that exists in feature subset selection as feature selection is initial step for determining the important attributes.

Kononenko, I. et. al [11] proposed an investigation based on theoretical and experimental analysis on mining of data set on the basis of weight, attribute rank etc. The irrelevant and redundant attributes that are removed in a way so that privacy can be preserved and data utility is maintained by determining the important feature of attribute.

III. Proposed Work

The proposed work below uses the united algorithm of k-anonymization, relieff (Feature subset selection), Association rule mining. The Proposed method has various advantages:

- It preserves the private data from being accessed.
- It maintains the data quality.
- It provides the data with no information loss.

Therefore, the proposed approach works in two phase: the first phase is the flowchart of obtaining a sanitized data and second phase is the algorithm that cleans the noisy data and generates the results related to minimum support and time for the given datasets. (German credit and Titanic data set)

1) Proposed Flowchart

This diagram introduces a multiple level filtering techniques with the help of combining various algorithms such as anonymization, relieff, and association rule mining. The datasets are collected from different sites to generate the results. The given two data sets for the aggregated process are "German credit data set" and "Titanic data set." The filtering of data is done at three levels by:

1. Frequent item set mining.
2. Column based filtering with the help of the relieff algorithm.
3. Row based filtering with the help of association rule mining.

Finally, in the end, the sanitized data is provided for both the data sets with the number of rules mined. It shows the graph between the time complexity and the minimum support and provides the comparison between the time taken by both the datasets to generate the sanitized database. So, that it can be analyzed the exact amount of

data that has been secured when initiated in the whole process.

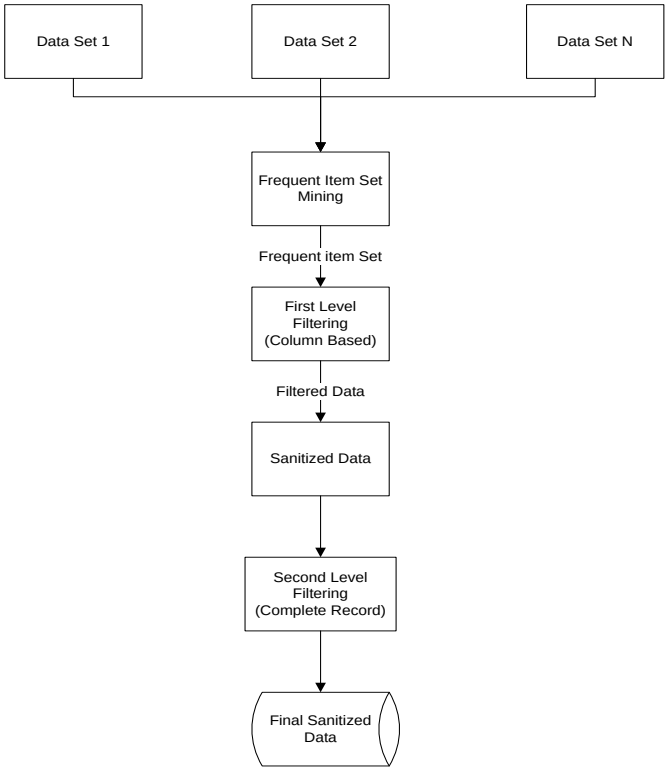


Fig 2. Flowchart of proposed work

2) Proposed Algorithm

Input: data sets D, minsup (Minimum Support Threshold), minconf (Minimum Confidence Threshold)

Output: The Filtered Dataset

Method:

1. BEGIN
2. For Each Dataset Di do
3. K-Data = Kanonymity(Di)
4. Selected Features Set F= Reflieff(K-Data)
5. For each Feature *f* in Set F
   Column-Filter (K-Data) → CF-Data
   End
6. Using MinSup and MinConf mine Privacy Breaching rules R using Association Rule Mining
7. For Each Rule *r* in set R
   Row-Filter (CF-Data) → Final-Data
   End
8. Return Sanitized Data
   End
9. Stop

IV. Experimental Results

The preliminary analysis is performed for given two data sets i.e. the " German credit data set" and "Titanic data set" at a minsup=0.02 and minconf=0.01 using the MATLAB tool and generate the results as given below for both the data sets.

	1 over_draft	2 credit_usage	3 credit_history	4 purpose	5 current_balance	6 Average_Credit_Balance	7 employment	8 location	9 personal_status	10 other_parties
1	'<0'	6	'critical/other e...	'radio/tv'	1169	'no known savings'	'>=7'	4	'male single'	'none'
2	'0<=X<200'	48	'existing paid'	'radio/tv'	5951	'<100'	'1<=X<4'	2	'female div/dep/...	'none'
3	'no checkin...	12	'critical/other e...	'education'	2096	'<100'	'4<=X<7'	2	'male single'	'none'
4	'<0'	42	'existing paid'	'furniture/e...	7882	'<100'	'4<=X<7'	2	'male single'	'guarantor'
5	'<0'	24	'delayed previo...	'new car'	4870	'<100'	'1<=X<4'	3	'male single'	'none'
6	'no checkin...	36	'existing paid'	'education'	9055	'no known savings'	'1<=X<4'	2	'male single'	'none'
7	'no checkin...	24	'existing paid'	'furniture/e...	2835	'500<=X<1000'	'>=7'	3	'male single'	'none'
8	'0<=X<200'	36	'existing paid'	'used car'	6948	'<100'	'1<=X<4'	2	'male single'	'none'
9	'no checkin...	12	'existing paid'	'radio/tv'	3059	'>=1000'	'4<=X<7'	2	'male div/sep'	'none'
10	'0<=X<200'	30	'critical/other e...	'new car'	5234	'<100'	'unemployed'	4	'male mar/wid'	'none'
11	'0<=X<200'	12	'existing paid'	'new car'	1295	'<100'	'<1'	3	'female div/dep/...	'none'
12	'<0'	48	'existing paid'	'business'	4308	'<100'	'<1'	3	'female div/dep/...	'none'
13	'0<=X<200'	12	'existing paid'	'radio/tv'	1567	'<100'	'1<=X<4'	1	'female div/dep/...	'none'
14	'<0'	24	'critical/other e...	'new car'	1199	'<100'	'>=7'	4	'male single'	'none'
15	'<0'	15	'existing paid'	'new car'	1403	'<100'	'1<=X<4'	2	'female div/dep/...	'none'
16	'<0'	24	'existing paid'	'radio/tv'	1282	'100<=X<500'	'1<=X<4'	4	'female div/dep/...	'none'
17	'no checkin...	24	'critical/other e...	'radio/tv'	2424	'no known savings'	'>=7'	4	'male single'	'none'
18	'<0'	30	'no credits/all p...	'business'	8072	'no known savings'	'<1'	2	'male single'	'none'
19	'0<=X<200'	24	'existing paid'	'used car'	12579	'<100'	'>=7'	4	'female div/dep/...	'none'
20	'no checkin...	24	'existing paid'	'radio/tv'	3430	'500<=X<1000'	'>=7'	3	'male single'	'none'
21	'no checkin...	9	'critical/other e...	'new car'	2134	'<100'	'1<=X<4'	4	'male single'	'none'
22	'<0'	6	'existing paid'	'radio/tv'	2647	'500<=X<1000'	'1<=X<4'	2	'male single'	'none'
23	'<0'	10	'critical/other e...	'new car'	2241	'<100'	'<1'	1	'male single'	'none'
24	'0<=X<200'	12	'critical/other e...	'used car'	1804	'100<=X<500'	'<1'	3	'male single'	'none'

Fig 3. Unclean and un-sanitized German credit data set imported as table in MATLAB.

	1 over_draft	2 credit_usage	3 credit_history	4 purpose	5 current_balance	6 Average_Credit_Balance	7 employment	8 location	9 personal_status	10 other_parties
1	3	3	2	8	143	1	4	4	4	3
2	2	30	4	8	771	4	1	2	1	3
3	1	9	2	5	391	4	2	2	4	3
4	3	27	4	6	849	4	2	2	4	2
5	3	18	3	2	735	4	1	3	4	3
6	1	24	4	5	870	1	1	2	4	3
7	1	18	4	6	534	3	4	3	4	3
8	2	24	4	3	814	4	1	2	4	3
9	1	9	4	8	563	5	2	2	2	3
10	2	22	2	2	748	4	5	4	3	3
11	2	9	4	2	191	4	3	3	1	3
12	3	30	4	4	701	4	3	3	1	3
13	2	9	4	8	288	4	1	1	1	3
14	3	18	2	2	150	4	4	4	4	3
15	3	12	4	2	233	4	1	2	1	3
16	3	18	4	8	184	2	1	4	1	3
17	1	18	2	8	465	1	4	4	4	3
18	3	22	5	4	853	1	3	2	4	3
19	2	18	4	3	904	4	4	4	1	3
20	1	18	4	8	609	3	4	3	4	3
21	1	6	2	2	402	4	1	4	4	3
22	3	3	4	8	499	3	1	2	4	3
23	3	7	2	2	423	4	3	1	4	3
24	2	9	2	3	323	2	3	3	4	3

Fig 4. Clean and sanitized German credit data set.

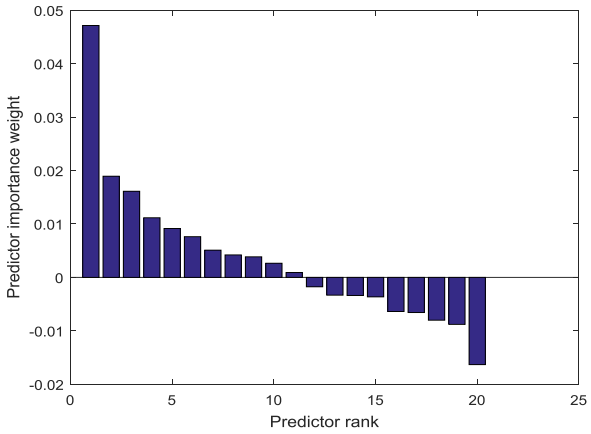


Fig 5. Feature subset selection procedure using Relieff algorithm

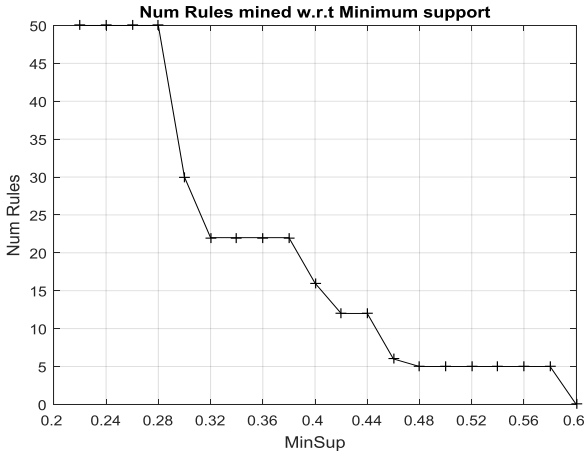


Fig 6. Number of rules mined according to minsup=0.2 and minconf=0.1

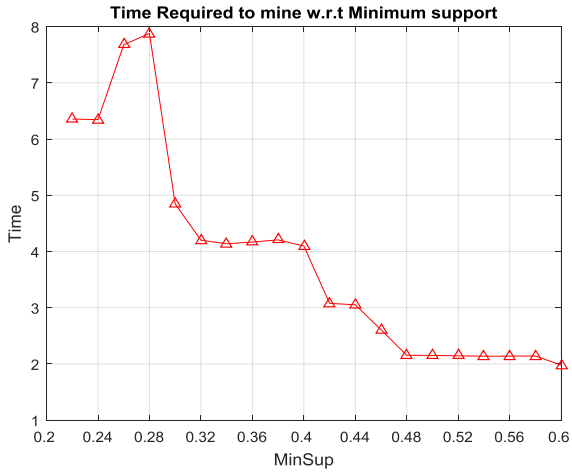
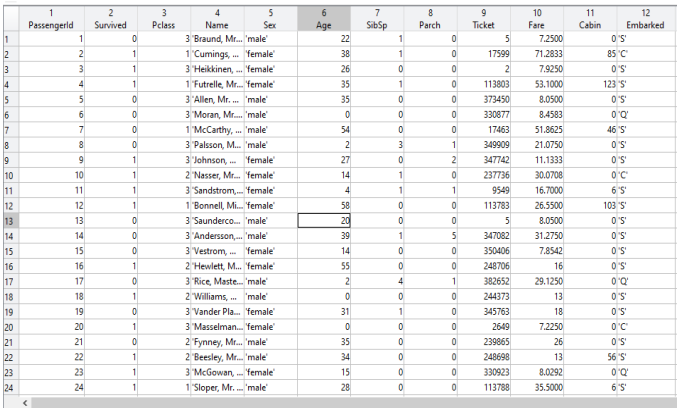
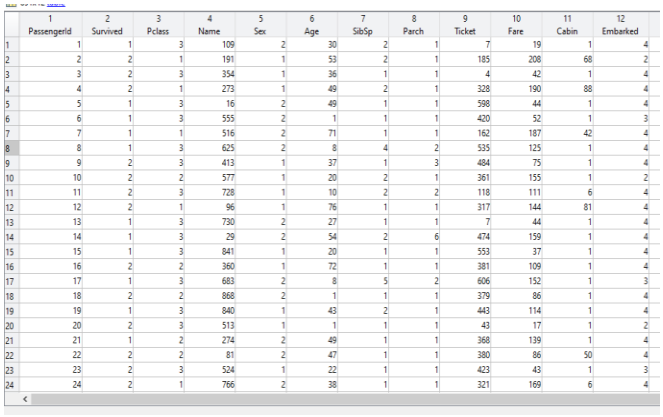


Fig 7. Time required to mine the data set



	1 PassengerId	2 Survived	3 Pclass	4 Name	5 Sex	6 Age	7 SibSp	8 Parch	9 Ticket	10 Fare	11 Cabin	12 Embarked
1	1	0	3	'Braund, Mr...	'male'	22	1	0	5	7.2500	0	'S'
2	2	1	1	'Cumings, ...	'female'	38	1	0	17599	71.2833	85	'C'
3	3	1	3	'Heikkinen, ...	'female'	26	0	0	2	7.9250	0	'S'
4	4	1	1	'Futrelle, Mr...	'female'	35	1	0	113803	53.1000	123	'S'
5	5	0	3	'Allen, Mr. ...	'male'	35	0	0	373450	8.0500	0	'S'
6	6	0	3	'Moran, Mr....	'male'	0	0	0	330877	8.4583	0	'Q'
7	7	0	1	'McCarthy, ...	'male'	54	0	0	17463	51.8625	46	'S'
8	8	0	3	'Palsson, M...	'male'	2	3	1	349909	21.0750	0	'S'
9	9	1	3	'Johnson, ...	'female'	27	0	2	347742	11.1333	0	'S'
10	10	1	2	'Nasser, Mr...	'female'	14	1	0	237736	30.0708	0	'C'
11	11	1	3	'Sandstrom,...	'female'	4	1	1	9549	16.7000	6	'S'
12	12	1	1	'Bonnell, Mi...	'female'	58	0	0	113783	26.5500	103	'S'
13	13	0	3	'Saunderco...	'male'	20	0	0	5	8.0500	0	'S'
14	14	0	3	'Andersson,...	'male'	39	1	5	347082	31.2750	0	'S'
15	15	0	3	'Vestrom, ...	'female'	14	0	0	350406	7.8542	0	'S'
16	16	1	2	'Hewlett, M...	'female'	55	0	0	248706	16	0	'S'
17	17	0	3	'Rice, Maste...	'male'	2	4	1	382652	29.1250	0	'Q'
18	18	1	2	'Williams, ...	'male'	0	0	0	244373	13	0	'S'
19	19	0	3	'Vander Pla...	'female'	31	1	0	345763	18	0	'S'
20	20	1	3	'Masselman...	'female'	0	0	0	2649	7.2250	0	'C'
21	21	0	2	'Fynney, Mr...	'male'	35	0	0	239865	26	0	'S'
22	22	1	2	'Beesley, Mr...	'male'	34	0	0	248698	13	56	'S'
23	23	1	3	'McGowan, ...	'female'	15	0	0	330923	8.0292	0	'Q'
24	24	1	1	'Sloper, Mr. ...	'male'	28	0	0	113788	35.5000	6	'S'

Fig 8. Unclean and un-sanitized Titanic data set as imported in table in MATLAB.



	1 PassengerId	2 Survived	3 Pclass	4 Name	5 Sex	6 Age	7 SibSp	8 Parch	9 Ticket	10 Fare	11 Cabin	12 Embarked
1	1	1	3	109	2	30	2	1	7	19	1	4
2	2	2	1	191	1	53	2	1	185	208	68	2
3	3	2	3	354	1	36	1	1	4	42	1	4
4	4	2	1	273	1	49	2	1	328	190	88	4
5	5	1	3	16	2	49	1	1	598	44	1	4
6	6	1	3	555	2	1	1	1	420	52	1	3
7	7	1	1	516	2	71	1	1	162	187	42	4
8	8	1	3	625	2	8	4	2	535	125	1	4
9	9	2	3	413	1	37	1	3	484	75	1	4
10	10	2	2	577	1	20	2	1	361	155	1	2
11	11	2	3	728	1	10	2	2	118	111	6	4
12	12	2	1	96	1	76	1	1	317	144	81	4
13	13	1	3	730	2	27	1	1	7	44	1	4
14	14	1	3	29	2	54	2	6	474	159	1	4
15	15	1	3	841	1	20	1	1	553	37	1	4
16	16	2	2	360	1	72	1	1	381	109	1	4
17	17	1	3	683	2	8	5	2	606	152	1	3
18	18	2	2	868	2	1	1	1	379	86	1	4
19	19	1	3	840	1	43	2	1	443	114	1	4
20	20	2	3	513	1	1	1	1	43	17	1	2
21	21	1	2	274	2	49	1	1	368	139	1	4
22	22	2	2	81	2	47	1	1	380	86	50	4
23	23	2	3	524	1	22	1	1	423	43	1	3
24	24	2	1	766	2	38	1	1	321	169	6	4

Fig 9. Clean, sanitized Titanic data set

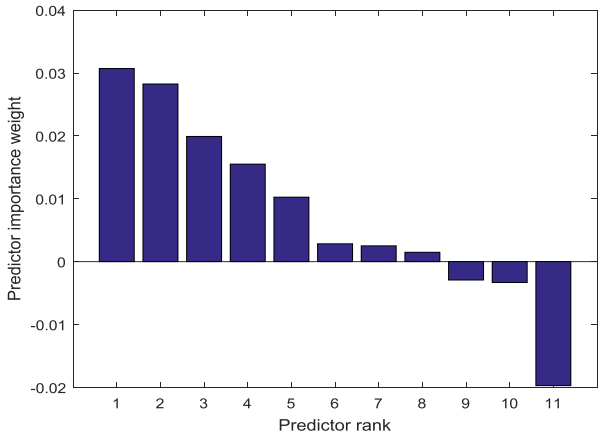


Fig 10. Feature subset selection procedure using Relieff algorithm.

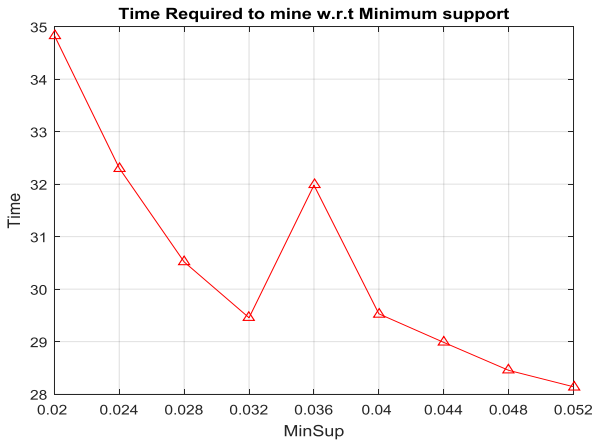


Fig 11. Time required to mine the data set

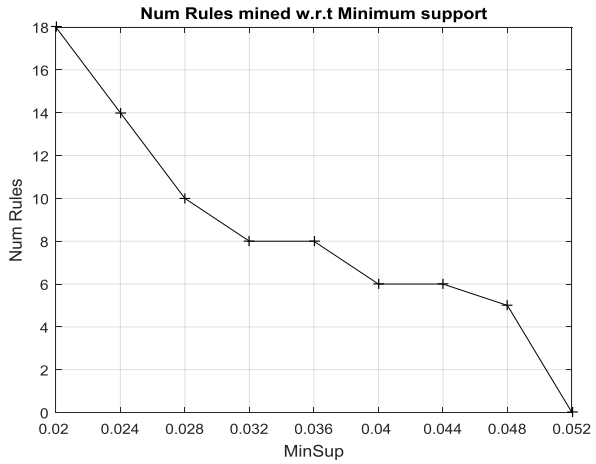
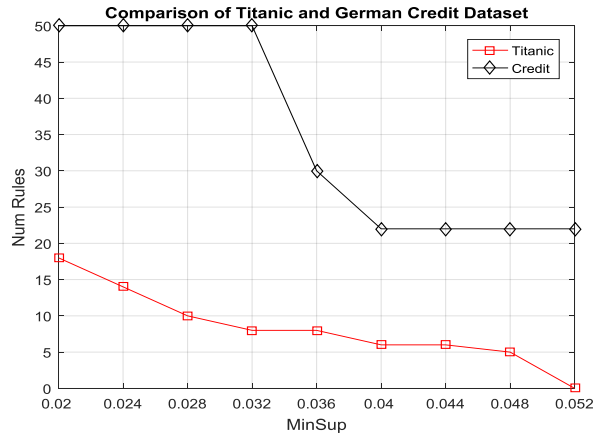


Fig 12. Number of rules mined according to minsup=0.02 and minconf=0.01



.

Fig 13. Comparison between the numbers of rules mined in both the data set.

Therefore, It can conclude that the result is shown in the figure no: 4 and 9 as shown above for the related two data sets i.e. German credit data set and Titanic data set. Both the data sets were worked under the combination of the algorithm as stated above and compute a novel filtering method for preserving the privacy of data in such a manner that minimizes the value of support, maximum is the privacy level and more rules are hidden for the given datasets. The figure no: 5 and 10 show the feature subset selection procedure using the relieff algorithm in MATLAB whereas the figure no: 6 and 12 demonstrate the result for some rules mined for the particular value of minimum support. Hence, the comparison between both the data sets regarding the number of rules extracted are shown in figure no: 13. Therefore, the final result can be concluded by the table given below i.e table 1.1. and 1.2.

Hence the database created is sanitized and this algorithm proves to be an effective method of securing privacy. Hence, the comparison states that the results for credit data set are better as compared to time taken and number of rules mined.

Table 1.1. Comparison between parameters of both the data sets

Characteristics of German credit data set and Titanic data set						
Data set	Min sup	Minconf	Time taken to mine data set		Rules mine	No.of iterations
			Total	Self		
German data set	0.2	0.1	92.430s	0.138s	50	20
Titanic data set	0.02	0.01	314.723s	0.134s	18	9

Table 1.2. Comparison of time taken in both the data sets.

German credit data set			
Function	Number of call	Total time	Self-time
German credit	1	92.430s	0.138s
Apriori	21	84.672s	9.308s

Anonymize	21	5.887s	3.386s
Relieff	1	1.086s	0.014s
Titanic data set			
Function	Number of call	Total time	Self-time
Titanic	1	314.723s	0.134s
Apriori	10	308.787s	37.583s
Anonymize	12	2.863s	1.616s
Relieff	1	0.713s	0.003s

V. Conclusion

Finally, from the above work, it can be concluded that the problem of privacy preservation caused due to the presence of noisy, un-sanitized data is somewhat resolved from the above stated algorithm based on anonymization, relieff and association rule mining. The proposed algorithm modifies the data according to the given value of minsup and minconf and initiates a new database i.e. Sanitized database for both the given datasets. The efficiency of both the datasets is shown in the above graphs. Hence, it can be demonstrated that the work proposed above is useful in maintaining the privacy to a certain level, hide more number of rules to secure data and this approach can be used in future to mine large set of traditional database.

References

[1] Aldeen, Y.A.A.S., Salleh, M. and Razzaque, M.A., 2015. A comprehensive review on privacy preserving data mining. SpringerPlus, 4(1), p.694.

[2] Patel, D.S., Tiwari, S. 2013. Privacy Preserving Data Mining. International Journal of Computer Science and Information Technologies, 4(1), pp.139-141.

[3] Vaidya, J. and Clifton, C., 2002, July. Privacy preserving association rule mining in vertically partitioned data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 639-644). ACM.

[4] Kantarcioglu, M. and Clifton, C., 2004. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE transactions on knowledge and data engineering, 16(9), pp.1026-1037.

[5] Mistry Nikita, J. and Kharwar Ankit, R., A Review on Association Rule Hiding.

[6] Gkoulalas-Divanis, A. and Verykios, V.S., 2010. Association Rule Hiding for Data Mining.

[7] G.Padmapriya. Hemalatha.M., 2014, December. Distributed Privacy preserving and Handling Privacy information leakage by using k-anonymity algorithm. In International Journal of Research in Applied Science & Engineering Technology (IJRASET).

[8] El Emam, K., and Dankar, F.K., 2008. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5), pp.627-637.

[9] Spolaôr, N., Cherman, E.A., Monard, M.C. and Lee, H.D., 2013, October. ReliefF for multi-label feature selection. In Intelligent Systems (BRACIS), 2013 Brazilian Conference on (pp. 6-11). IEEE.

[10] Durgabai, R.P.L., 2014. Feature selection using ReliefF algorithm. IJARCCE—International Journal of Advanced Research in Computer and Communication Engineering, 3(10).

[11] Robnik-Šikonja, M. and Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1-2), pp.23-69.