JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-4, Issue-7 E-ISSN: 2347-2693

Mining Association rules and Differential Privacy Preservation using Randomization

Krishna Kumar Tripathi
Sant Gadge Baba Amravati University,
Amravati, India

Dr. Narendra S. Chaudhari
Computer Science & Engineering
V.N.I.T Nagpur, India

Available online at: www.ijcseonline.org

Received:18/Jun/2016 Revised: 26/Jun/2016 Accepted: 16/Jul/2016 Published: 31/Jul/2016

Abstract – Paper herewith proposes an optimal predictive class association rule mining techniques for extracting the minimum rule having same predictive power of complete predictive class association rule by using predictive association rule set instead of complete class association rule , proposed methodologies in this paper can avoid the redundant and non-useful computation that would otherwise be required or needed for the mining of predictive class association rules and therefore improving the efficiency and effectiveness of the mining process significantly. Paper herewith presents an efficient and effective algorithm framework for mining the optimal predictive class association rule dataset by using CPAR before they are actually generated. In this paper, techniques have been implemented and obtained experimental results demonstrate that the algorithm generates the optimal class association rule set. Hence paper herewith propose a new data classification approach, Classification based on the Predictive Association Rules, which mainly combines the advantages and knowledge of both traditional rule-based and associative classification. Instead of generating the large number of candidate class association rules as in associative classification techniques, CPAR usually adopts a greedy algorithm for generating rules directly from the training dataset.

Keywords –Discrimination, Association, CPAR, GC, DDPD, DDPP, IDPD, IDPP, DRP, IRP

I. INTRODUCTION

The main goal of the predictive class association rule mining is to identify and extract all the rules which are satisfying some of the basic criteria and requirement, such as minimum support threshold and minimum confidence threshold. It was initially suggested for solving the common market basket problem observed in the transaction databases, and has then extended to analysis and solves many other problems such as classification problem. A set of predictive class association rules for the requirement of classification is called predictive association rule set. Mainly, predictive association class rules are dependent on the relational databases, and hence the consequences of rules are represented in the pre-specified columns of the tables, class (decision) attribute. Clearly, in this situation a relational database can be then mapped and transformed to the transaction database by taking consideration of attribute and attribute value pair as an item set. After having being mapped into a relational database into an transaction database, a decision class predictive class association rule set , which is basically an subset of predictive association rules with the specified set of the classes as their consequences, and hence a predictive class association rule set is a small subset of derived class association rule set [1]. (Classification based on the Predictive Association Rules). CPAR inherits and represents the basic property of the FOIL methodologies [2] in rule generation approach and combines the features of class associative classification rules in predictive rule analysis techniques. In comparison with the associative classification, CPAR has the following advantages: (1) CPAR algorithm usually generates a much smaller subset of high-quality predictive class associations rules inherited directly from the training dataset; (2) to avoid generating redundant and non-useful rules, CPAR algorithm generates each predictive association rule by considering the set of “already generated” class association rules; and (3) In the prediction of class label of an example, CPAR generally identifies the best k possible rules such that this mentioned example criteria satisfies. CPAR generates a smaller set of predictive class association rules, with lower redundancy and high quality in comparison with predictive class associative classification.

Hence, CPAR algorithm is more accurate, effective and time-efficient in the both prediction and rule generation and also achieves as high accuracy as of the predictive associative classification techniques.

Section (ii) of this paper deals with the literature survey, Section (iii) focuses on methodoly functioning, section (iv) deals with experimental results while section (v) concentrate on the performance measures of the system. Section (vi) refers to the various utility measure used for performance tracking and section (vii), (viii) and (ix) focuses on conclusion, future work and references respectively.

[*] Corresponding Author: K K Tripathi
e-mail: tripathi_kk@gmail.com

1.1 Privacy preservation goals

Firstly, comprehensive study on the aspects of mining predictive association rules and privacy preservation using randomization is conducted in order address the issue of detrimental treatment and identification of the people based on the sensitive attributes, no major work with proper mining predictive association rule and clustering is carried out in this area and hence generated the motivation to carry out the research and implementation in this area. Secondly research proved that the existing techniques do not have enhanced mining predictive association rule techniques, clustering & privacy preservation mechanisms [3].

The prime objective and requirement of this paper is to develop an web based framework and architecture which can provide discrimination prevention, mining predictive association rule, perform clustering and classification as well as privacy preservation of the data .With the above proposed work we are planning to achieve the following goals [4].

- Improvement in the existing discrimination approach.
- Obtaining better results for discrimination prevention and privacy preservation.
- Obtaining better results for privacy preservation.
- Creating a new models and architecture for the discrimination prevention.
- Clustering of the training data set.
- Finding predictive association rules present in the training data set.
- Differential privacy preservation using randomization approach.
- Comparison of the proposed architecture with the existing methods and techniques.
- Utility measures for degree of discrimination and privacy preservation.

II. PREVIOUS WORK DONE

Mining predictive class association rules in the transactional data set [5] is a central task of data mining and has really shown applications and demand in various research areas [6][7][8]. Currently most of proposed algorithms methodologies for mining predictive class association rules are dependent on the mining of the dataset by using Apriori classification techniques [9], and used the approach so-called as the 'downward closure' property which mentions that all the subsets of an frequent training data item set must be frequent in nature. Example of these techniques can be found in the References. [10][11]. A symmetric expression of the mentioned downward closure property is basically an upward closure property of all the supersets training item set of an infrequent class item set must be infrequent in nature. Finding predictive class classification rules has been an important and new research focus area in the aspect of the machine learning communities [12][13]. Mining predictive class classification rules can be viewed as special types of the mining class predictive association rules, since here a set of predictive association rules with pre-identified (classes) consequences can be considered for the classification. Methodologies for mining predictive class association rules have already being taken into consideration of mining classification rules.

Particularly, results mentioned in are partially encouraging, since it can build more accurate and effective classifiers than those from C4.5. But, the methods are not very effective and efficient because it uses Apriori-based algorithm to generate the (Decision) class association rules that can be very large in number when the value of minimum support is small. Generally saying, predictive class association rule set is basically a type of target-constraint based classification association rules. Interesting rule sets and Constraint rule sets belongs to this type of rule set. Main Problems associated with this type of rule sets are that they either exclude or miss some useful predictive class association rules, or they contain many of the redundant non useful rules which are of no use for the prediction of the behavior and decision support system. Along with this, methodology used for mining these rule sets handle only one type of target at given point of time (basically building of one enumeration tree to identify), hence these approach cannot be effectively and efficiently used for mining predictive class association rules that are on based on the multiple classes, mainly when the total number of classes is large in item set. Our optimal predictive class association rule set differs from these mentioned rule sets as it's minimal in size and keeps track all the predictive power.

2.1 Analysis of the problem

During the study analysis, investigation and identification of the above literature survey and previous work, we have come across some of the issues and limitations that were explored and are summarized below:

- The relationship between differential privacy preservation approaches and discrimination prevention methodologies in mining of the dataset is not investigated, identified and researched. It always remains untouched topic for analyzers' and research to identify whether differential privacy preservation mechanism can help in the area of the anti-discrimination or vice-versa.
- The methodology focuses on the methods to find out the discrimination in the original training data only for one of the discriminatory item and also it is based on a single measuring approach.

- They usually doesn't include any techniques , methods and framework to identify and evaluate that what amount of discrimination has been removed from the training dataset and the what is the total amount of information loss has been incurred due to above approach.
- The synergies among rule hiding techniques for discrimination removal and rule hiding methods in differential privacy preserving data mining is not evaluated and published.
- They mainly focused on either direct discrimination prevention or indirect discrimination or not on both at the same time.
- The techniques also doesn't show improved methods for mining predictive association rule , perform clustering and classification techniques

2.1 Privacy Aware Data Mining Process

Privacy preservation in data mining is basically not just a goal or service like security, but it is the belief of the user's to reach a protected and controllable state, mainly even without having to actively monitor for it by themselves. Hence, privacy preservation is described as "the rights of the individual's or people to identify for themselves when, what and how information about them is used for various goals, reason and purpose". The preservation and protection of responsive training data is an essential approach of research which has involved many of the researchers in field of information technology privacy preservation. In the discovery of and attempt at the assuring privacy when sharing and mining personal individual's data have led to introduction of the privacy preserving in data mining (PPDM) methodologies and approaches. [4]

III. METHODOLOGY FOR FUNCTIONING

Services offered by the proposed web-based systems are preventing discrimination, clustering, classification, finding predictive association rules and randomization approach of privacy preservation:

1. Clustering of the training dataset.
2. Mining frequent predictive association rules from the preprocessing data set.
3. Randomization approach for Differential privacy preservation.
4. Training Dataset Post processing.
5. Storing the discretized preprocessed, and post processed data in to the database.
6. Reporting and analytics section.
7. Experimental results.
8. Performance utility measure.

German Credit dataset that can be obtained from ftp.ics.uci.edu/pub/machine-learning-databases/statlog/ location is used for the proposed thesis work. German Credit data set totally contains 1000 records, 13 nominal and 7 numeric attributes, with credit as a class label that can be good or bad.

IV. EXPERIMENT RESULT

The operation and functioning ability of each of the service present in the web based framework is discussed in details in the following section.

1. Loading data module will browse the training data set and upload that to the framework for processing.
2. Clustering is a mechanism of partitioning a set of objects /data (or) into a subset of meaningful classes and sub-classes, called clusters. Finds natural order grouping of instances given based on the un-labeled data, it helps viewers to understand the natural structure or grouping/ ordering in a data set.

   K means Algorithm is used for clustering.

   In figure 1, Clustering is performed by using following algorithm

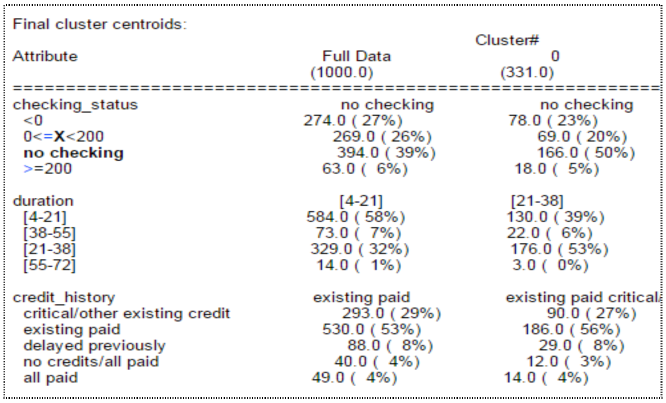
Let X = {x1,x2,x3,……..,xn} be set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers from the training data.
2) Calculate the distance between each of the data point from the cluster centers.
3) Assign the data point to cluster center whose distance from the center of the cluster is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$(x + a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

$$V_i = (1/C_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' represents the number of data points in ith cluster.
5) Recalculate the distance between each of the data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise go to step and repeat.



```
Final cluster centroids:
                                                    Cluster#
Attribute                        Full Data                 0
                                  (1000.0)           (331.0)
==============================================================
checking_status                 no checking        no checking
  <0                           274.0 ( 27%)       78.0 ( 23%)
  0<=X<200                       269.0 ( 26%)       69.0 ( 20%)
  no checking                    394.0 ( 39%)      166.0 ( 50%)
  >=200                         63.0 (  6%)        18.0 (  5%)

duration                          [4-21]           [21-38]
  [4-21]                       584.0 ( 58%)      130.0 ( 39%)
  [38-55]                       73.0 (  7%)       22.0 (  6%)
  [21-38]                      329.0 ( 32%)      176.0 ( 53%)
  [55-72]                       14.0 (  1%)        3.0 (  0%)

credit_history                 existing paid     existing paid critical
  critical/other existing credit   293.0 ( 29%)       90.0 ( 27%)
  existing paid                530.0 ( 53%)      186.0 ( 56%)
  delayed previously              88.0 (  8%)       29.0 (  8%)
  no credits/all paid            40.0 (  4%)       12.0 (  3%)
  all paid                     49.0 (  4%)       14.0 (  4%)
```

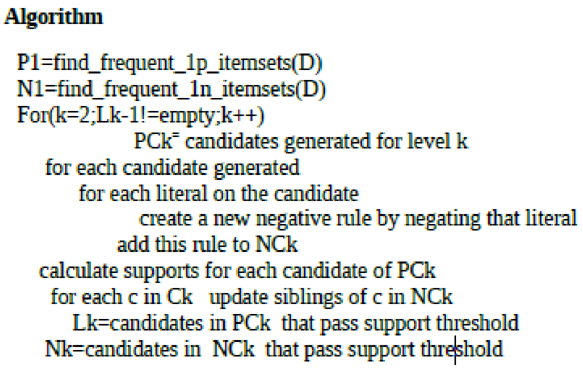
Figure 1. Clustering data set

3. Predictive Association rules are created by analyzing the training data in order to find the frequent if/then patterns and then using the criteria of support and confidence to identify the most important and common relationships.

Associative classification: Major steps

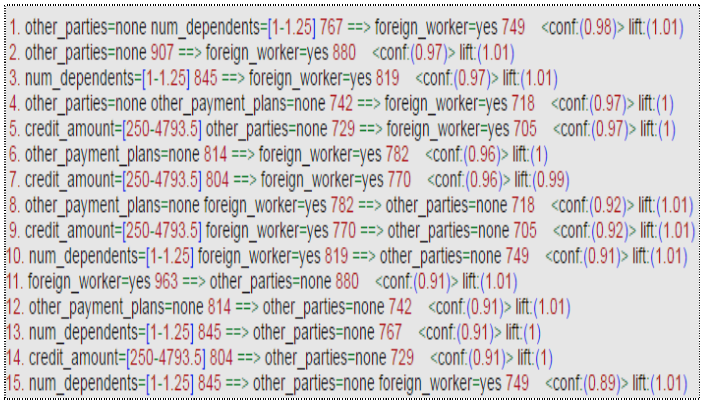
- Mine the training dataset to find strong predictive associations rules between frequent patterns (conjunctions of attribute-value pairs) and association rule class labels.
- Association rules are basically represented in the form of $P_1 \wedge p_2 \ldots \wedge p_l \rightarrow$ "$A_{class} = C$" (conf, sup)
- Organize the association rules so that they form a rule-based classifier.

In figure 2, Classification is performed by using following algorithm,



Algorithm

```
P1=find_frequent_1p_itemsets(D)
N1=find_frequent_1n_itemsets(D)
For(k=2;Lk-1!=empty;k++)
                PCk- candidates generated for level k
    for each candidate generated
         for each literal on the candidate
                  create a new negative rule by negating that literal
              add this rule to NCk
   calculate supports for each candidate of PCk
    for each c in Ck   update siblings of c in NCk
       Lk=candidates in PCk  that pass support threshold
   Nk=candidates in  NCk  that pass support threshold
```



```
1. other_parties=none num_dependents=[1-1.25] 767 ==> foreign_worker=yes 749    <conf:(0.98)> lift:(1.01)
2. other_parties=none 907 ==> foreign_worker=yes 880    <conf:(0.97)> lift:(1.01)
3. num_dependents=[1-1.25] 845 ==> foreign_worker=yes 819    <conf:(0.97)> lift:(1.01)
4. other_parties=none other_payment_plans=none 742 ==> foreign_worker=yes 718    <conf:(0.97)> lift:(1)
5. credit_amount=[250-4793.5] other_parties=none 729 ==> foreign_worker=yes 705    <conf:(0.97)> lift:(1)
6. other_payment_plans=none 814 ==> foreign_worker=yes 782    <conf:(0.96)> lift:(1)
7. credit_amount=[250-4793.5] 804 ==> foreign_worker=yes 770    <conf:(0.96)> lift:(0.99)
8. other_payment_plans=none foreign_worker=yes 782 ==> other_parties=none 718    <conf:(0.92)> lift:(1.01)
9. credit_amount=[250-4793.5] foreign_worker=yes 770 ==> other_parties=none 705    <conf:(0.92)> lift:(1.01)
10. num_dependents=[1-1.25] foreign_worker=yes 819 ==> other_parties=none 749    <conf:(0.91)> lift:(1.01)
11. foreign_worker=yes 963 ==> other_parties=none 880    <conf:(0.91)> lift:(1.01)
12. other_payment_plans=none 814 ==> other_parties=none 742    <conf:(0.91)> lift:(1.01)
13. num_dependents=[1-1.25] 845 ==> other_parties=none 767    <conf:(0.91)> lift:(1)
14. credit_amount=[250-4793.5] 804 ==> other_parties=none 729    <conf:(0.91)> lift:(1)
15. num_dependents=[1-1.25] 845 ==> other_parties=none foreign_worker=yes 749    <conf:(0.89)> lift:(1.01)
```

Figure 2. Association rules

Privacy preservation in data mining can be implemented in many ways and here we have done by use of randomization approach. Hide the original training data set by randomly changing the data values by using some of additive noise but still preserving the patterns and associations present among the original data (to preserve the underlying probabilistic data properties) and then reconstruct the distribution of the original training data values from the perturbed or modified data.

The main goals PPDM techniques are:

- PPDM methods should be capable to discover sensible information.
- It should be able to resistant the various data mining techniques.
- It should not be compromise the access and use of no sensitive data.
- It should not have exponential computational calculations and complexity

Numerical Randomization

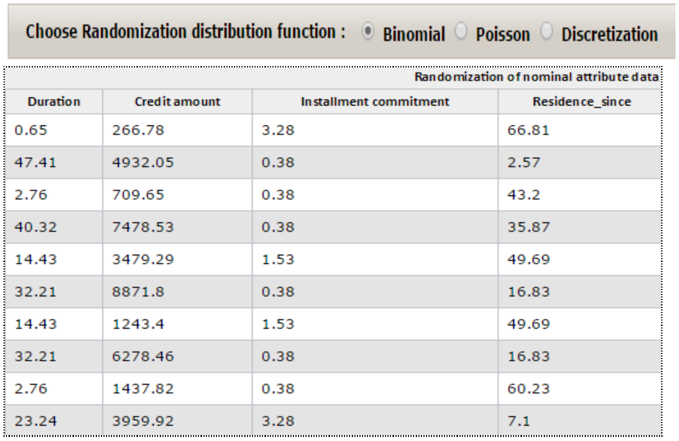
In Figure 3 and 4, Let each records Ri, i = 1, 2. . . N, have a numerical attribute xi. Let's assume that each of the xi is an instance of any random variable Xi, where all of the Xi are identically and independent distributed. The cumulative distribution function (CDF) (the same for every Xi) is denoted by the function FX. The server wants to identify and learn the function FX, or its closest approximation; this is an aggregate technique which the processing server is allowed to know. The server can get any information about the clients that can be derived from the model, but we would like to restricts / limit what the server knows about the actual records / instances xi to preserve the privacy of the data [14].

Each client randomizes user's records its xi by adding the random shift mi. The shifting values mi are independently and identically distributed random variables with cumulative distribution function (CDF) FM; their distribution is chosen well in advance and also known to the server. Thus, client Ci then sends randomized value zi = xi + mi to the processing server, and then the server's task is to identify the approximate function FX given FM and values z1, z2, . . . , zN [15].

This methodology attempts to hide the sensitive data by randomly modifying the data values often using additive noise.
The dataset returns a value 'u*v' where u is the original data, and v is a random value drawn from a certain distribution.
Most commonly used distributions are the Normal / Binomial and Poisson distribution and it will apply on the Numeric Attribute of the dataset.

Randomization can be performed in 3 ways

1) Randomization by Binomial distribution.
2) Randomization by Poisson's distribution.
3) Randomization by Discretization.

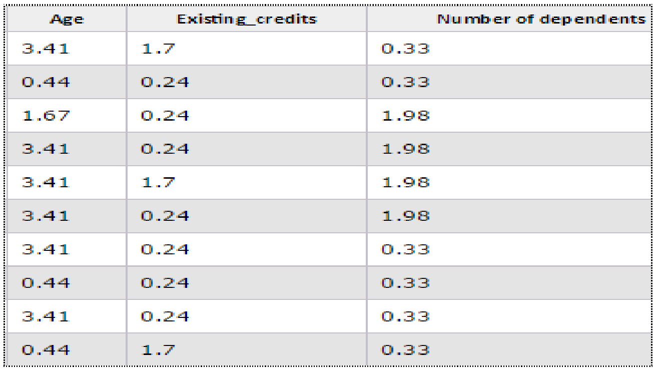


Choose Randomization distribution function : ● Binomial ○ Poisson ○ Discretization

Randomization of nominal attribute data

Duration	Credit amount	Installment commitment	Residence_since
0.65	266.78	3.28	66.81
47.41	4932.05	0.38	2.57
2.76	709.65	0.38	43.2
40.32	7478.53	0.38	35.87
14.43	3479.29	1.53	49.69
32.21	8871.8	0.38	16.83
14.43	1243.4	1.53	49.69
32.21	6278.46	0.38	16.83
2.76	1437.82	0.38	60.23
23.24	3959.92	3.28	7.1

Figure 3. Randomization



Age	Existing_credits	Number of dependents
3.41	1.7	0.33
0.44	0.24	0.33
1.67	0.24	1.98
3.41	0.24	1.98
3.41	1.7	1.98
3.41	0.24	1.98
3.41	0.24	0.33
0.44	0.24	0.33
3.41	0.24	0.33
0.44	1.7	0.33

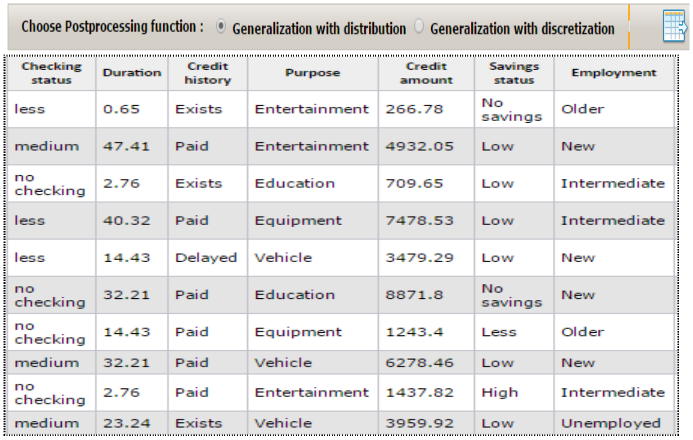
Figure 4. Randomization

4. After the data analysis, the viewers can perform some post-processing task on the results data, this post-processing facilitates the analysis of the previous result sets obtained in a formal way. The main objective is to extract and analysis the meaningful results to perform certain decision making results, It allows to modify the resulting data mining models, instead of totally cleaning the original training data set The post processing technique does not allow the whole data set to be published: but only the modified data mining models distributions can be published (knowledge publishing) for analysis

Figure 5 and 6 shows the Post processed data. Post processing can be performed by generalization or discretization.

1. Input: Credit data set file.
2. Read the record from input file.
3. If the value is nominal go to step 3 – else go to step 4 to 5.
4. Replace the nominal values with their generalized values
5. Calculate the equivalent normal distribution values for numeric data.
6. Replace the numeric values with the output of step 4.
7. Generate the new record based on the steps 3-6.
8. Repeat steps 2-6 for all the records.
9. End.

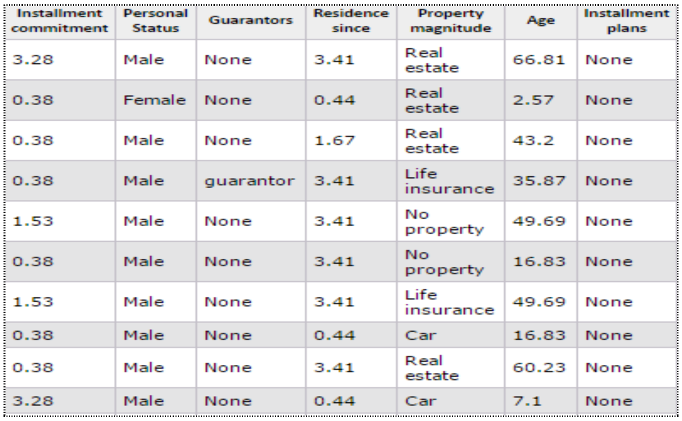
Post processing can be performed by generalization or discretization



Choose Postprocessing function : ● Generalization with distribution ○ Generalization with discretization

Checking status	Duration	Credit history	Purpose	Credit amount	Savings status	Employment
less	0.65	Exists	Entertainment	266.78	No savings	Older
medium	47.41	Paid	Entertainment	4932.05	Low	New
no checking	2.76	Exists	Education	709.65	Low	Intermediate
less	40.32	Paid	Equipment	7478.53	Low	Intermediate
less	14.43	Delayed	Vehicle	3479.29	Low	New
no checking	32.21	Paid	Education	8871.8	No savings	New
no checking	14.43	Paid	Equipment	1243.4	Less	Older
medium	32.21	Paid	Vehicle	6278.46	Low	New
no checking	2.76	Paid	Entertainment	1437.82	High	Intermediate
medium	23.24	Exists	Vehicle	3959.92	Low	Unemployed

Figure 5. Post processing



Installment commitment	Personal Status	Guarantors	Residence since	Property magnitude	Age	Installment plans
3.28	Male	None	3.41	Real estate	66.81	None
0.38	Female	None	0.44	Real estate	2.57	None
0.38	Male	None	1.67	Real estate	43.2	None
0.38	Male	guarantor	3.41	Life insurance	35.87	None
1.53	Male	None	3.41	No property	49.69	None
0.38	Male	None	3.41	No property	16.83	None
1.53	Male	None	3.41	Life insurance	49.69	None
0.38	Male	None	0.44	Car	16.83	None
0.38	Male	None	3.41	Real estate	60.23	None
3.28	Male	None	0.44	Car	7.1	None

Figure 6. Post processing

V. PERFORMANCE MEASURES

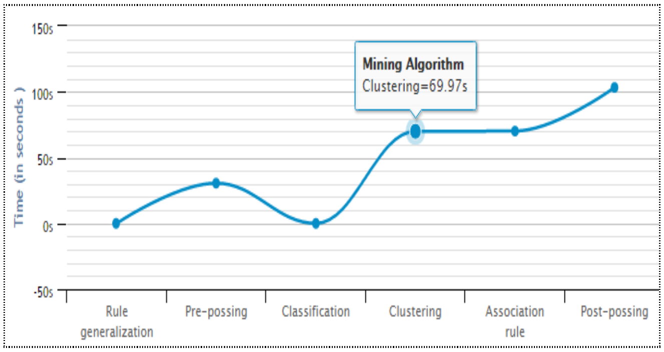


Figure 7. Execution time for mining algorithms

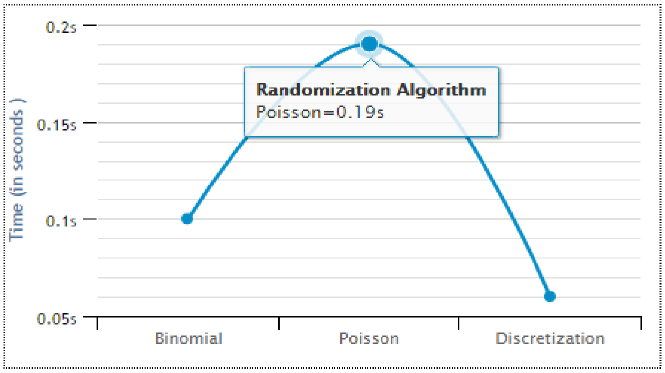
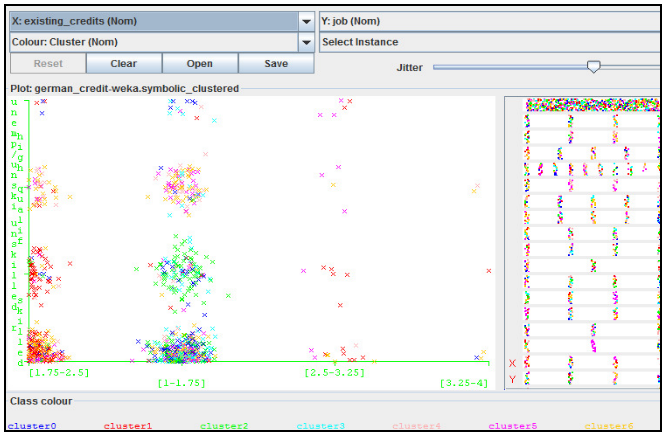


Figure 8. Execution time for randomization algorithms

Figure 7 shows the execution time for data mining algorithm , on the x – axis data mining algorithm are shown and on the y –axis , time taken by the algorithms (in seconds) are shown.

Figure 8 shows the execution time for randomization algorithm, on the x – axis randomization algorithm are shown and on the y –axis, time taken by the randomization algorithms (in seconds) are shown.



X: existing_credits (Nom)

Y: job (Nom)

Colour: Cluster (Nom)

Plot: german_credit-weka.symbolic_clustered

Figure 9. Visualize graph of cluster model (Job on Y and existing credits on X axis

VI. UTILITY MEASURES AND METHODS

Direct & Indirect Discrimination Prevention Method

M = {DRP, IRP}

DRP = Direct Rule Protection Technique :- The DRP technique changes the class item --> X to X' for some records that satisfy the condition, --> A, B --> X'.

IRP = Indirect Rule Protection Technique: - IRP technique changes the class item --> Y to Y' for some records that satisfy the condition, --> A, B, --> D --> Y'.

Performance Measures P = {DDPD, DDPP, IDPD, IDPP}

DDPD = Direct discrimination prevention degree calculates and quantifies the percentage of α-discriminatory rules which are no longer α-discriminatory in the transformed training data set.

DDPD = | CR |- | CR'| / | CR |

Where CR is the database of α-discriminatory rules that are extracted from DB and CR' is the database of α-discriminatory rules that are extracted from the transformed data set DB'

DDPP = Direct discrimination protection preservation measure It quantifies the percentage of α-protective rules in original dataset that remain α- protective in the transformed data set.

DDPP = | QR | ∩ | QR'| / | QR |

where QR is the database of α-protective rules that are extracted from the original data set DB and QR' is the database of α-protective rules that are extracted from the transformed data set DB'.

IDPD = Indirect discrimination prevention degree calculates and quantifies the percentage of redlining rules which are no longer redlining in the transformed data set. It is measured as DDPD but substituting CR and CR' with the database of redlining rules extracted from DB and DB', respectively

IDPP = Indirect discrimination protection preservation quantifies the percentage of the non-redlining rules in original data set that are remain non-redlining rules in the transformed data set. It is defined like DDPP but substituting QR and QR' with the database of nonredlining extracted from DB and DB', respectively.

Output O = {X'}

O is the set of the outputs from the system. X' is the set of the transformed dataset which is free from direct and indirect discrimination.

To measure the quality of the data, we have used two metrics proposed in the literature survey as the measures of information loss in the context of data hiding for privacy-preserving data mining (PPDM)

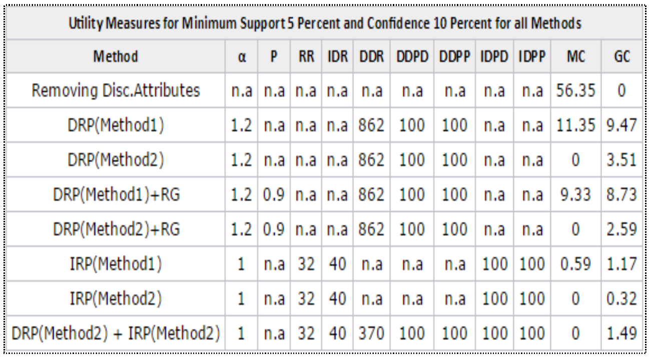
Misses cost (MC). This measure defines the percentage of rules among those which are extractable from the original training data set which cannot be extracted from the transformed data set (mainly because of side effect of the transformation process) [16].

Ghost cost (GC). This measure defines the percentage of rules among that are extractable from the transformed data set that were not extractable from the original training data set (mainly because side effect of the transformation process).

MC and GC should ideally be calculated as 0 percent. However, MC and GC may not be always 0 percent because of side effect of the transformation process.

Table 1 demonstrate the results for minimum support value 5 percent and minimum confidence value 10 percent. The results of the direct discrimination prevention techniques are reported for discriminatory threshold having α = 1.2 and, in the cases where direct rule protection technique is applied in combination with the rule generalization, we used p = 0:9, and DI(8) = {Foreign worker = Yes, Personal Status = Female and not Single, Age = Old} in the data set

In addition to the above results, the results of the indirect discrimination prevention techniques and both direct and indirect discrimination prevention methods are reported for discriminatory threshold α= 1 and DI(8) = {Foreign worker = Yes}.



Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for all Methods

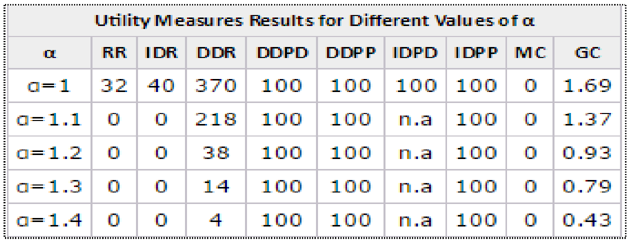
Method	α	P	RR	IDR	DDR	DDPD	DDPP	IDPD	IDPP	MC	GC
Removing Disc.Attributes	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	n.a	56.35	0
DRP(Method1)	1.2	n.a	n.a	n.a	862	100	100	n.a	n.a	11.35	9.47
DRP(Method2)	1.2	n.a	n.a	n.a	862	100	100	n.a	n.a	0	3.51
DRP(Method1)+RG	1.2	0.9	n.a	n.a	862	100	100	n.a	n.a	9.33	8.73
DRP(Method2)+RG	1.2	0.9	n.a	n.a	862	100	100	n.a	n.a	0	2.59
IRP(Method1)	1	n.a	32	40	n.a	n.a	n.a	100	100	0.59	1.17
IRP(Method2)	1	n.a	32	40	n.a	n.a	n.a	100	100	0	0.32
DRP(Method2) + IRP(Method2)	1	n.a	32	40	370	100	100	100	100	0	1.49

Table 1. results for support 5 and minimum confidence 10

α: discriminatory threshold α
P: confidence
RR: No of redlining rules
IDR: No of α indirect disc rules
DDR: No of α direct disc rules
Misses cost (MC): percentage of lost rules
Ghost cost (GC): percentage of introduced rules

Min support 5%, min confidence 10%
32340 frequent classification rules
22763 background knowledge rules
32 redlining rules
40 indirect rules
862 direct discriminations

In Table 2 shown below, the results are obtained for different values of α --> [1,1.4] . We selected these α intervals in such a manner that, with respect to the predetermined discriminatory items in this experiment for the German Credit card data set i.e DI(8) = {Foreign worker = Yes} both direct α-discriminatory and redlining rules could be extracted.



Utility Measures Results for Different Values of α

α	RR	IDR	DDR	DDPD	DDPP	IDPD	IDPP	MC	GC
α=1	32	40	370	100	100	100	100	0	1.69
α=1.1	0	0	218	100	100	n.a	100	0	1.37
α=1.2	0	0	38	100	100	n.a	100	0	0.93
α=1.3	0	0	14	100	100	n.a	100	0	0.79
α=1.4	0	0	4	100	100	n.a	100	0	0.43

Table 2. results for different values of α --> [1,1.4]

Above table demonstrate that the proposed solution achieves a high degree of the discriminatory threshold for both direct and indirect discrimination prevention for different values. The main important point here is that, by applying the proposed technique, we get good results for both direct and indirect discrimination prevention at the same time and In addition, the values of GC and MC demonstrate that the proposed techniques incurs low information loss of the training dataset

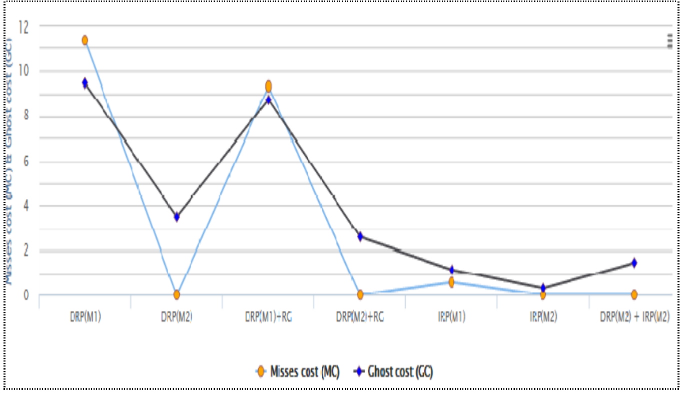


Figure 10. German Credit dataset: Utility methods Vs MC and GC

This elift states being foreign worker increases the probability and chances of denying credit w.r.t. all other people who have denied credit and who have asked credit for radio/TV. [Purpose=Radio/TV] is called context of the rule.

α is a fixed discriminatory threshold, which states an acceptable level of discrimination according to rules and regulations. The PD classification rule is called α-protective if elift < α and if elift >= α, the rule is called α-discriminatory.

The above rule is α-discriminatory as elift > α. Given α = 1. Application discovers such kind of α-discriminatory rules in different contexts, so the german credit dataset is discriminatory w.r.t. Foreign Worker.

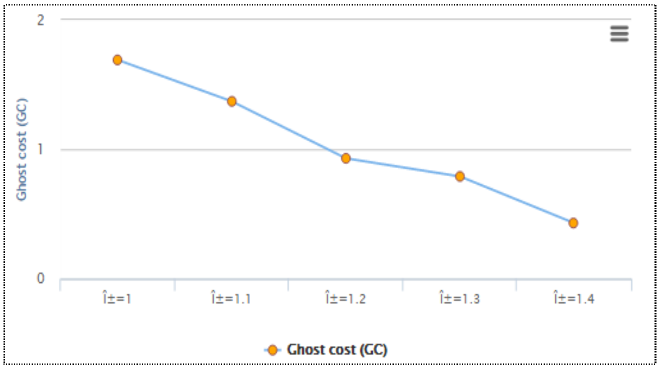


Figure 11. German Credit dataset: α Vs GC

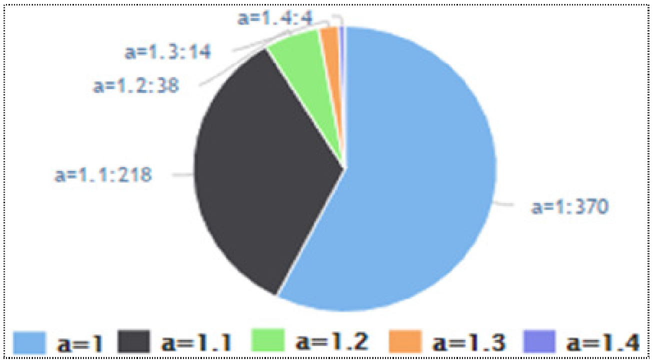
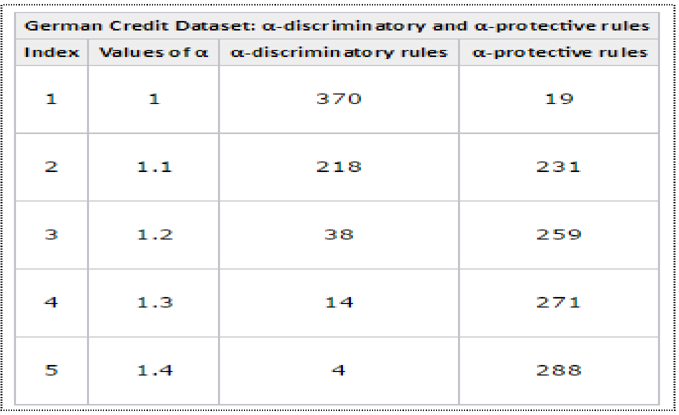


Figure 12. No of α direct disc rules

Table 3 depicts values of number of the α-discriminatory and α-protective rules for the different values of α for German credit training dataset.

Graph in figure 11 , represents variations in the values of α-discriminatory and α-protective rules with different the values of α. X-axis represents different values of α and Y-axis represents number of α-discriminatory and α-protective rules.

From the graph, it can be concluded that, as the value of α increases, number of the α-discriminatory rules decreases and number of the α-protective rules increases. That is discrimination reduces as we increase the value of α.



German Credit Dataset: α-discriminatory and α-protective rules

Index	Values of α	α-discriminatory rules	α-protective rules
1	1	370	19
2	1.1	218	231
3	1.2	38	259
4	1.3	14	271
5	1.4	4	288

Table 3. number of the α-discriminatory and α-protective rules

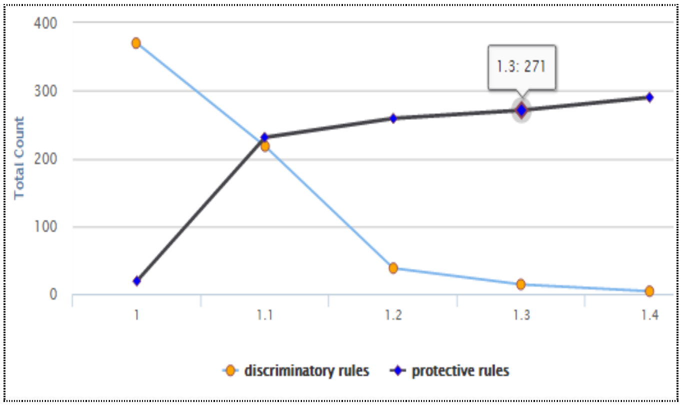


Figure 13. α versus α-discriminatory and α-protective rules

VII. CONCLUSION

The above solution proposed is the web-based framework for mining predictive association rules and privacy preservation using randomization.

It can be concluded from the graph that, as we increase the value of α, number of the α-discriminatory rules decreases and number of the α-protective rules increases. I.e. discrimination reduces as value of α increases. Here registered user can perform Clustering of the training dataset ,Mining frequent predictive association rules from the preprocessing data set , use randomization approach for Differential privacy preservation ,Training Dataset Post processing , can store the discretized , preprocessed, and post processed data in to the database and reporting and analytics section

VIII. FUTURE WORK

Future work will emphasize on the elaborating how the proposed web based framework can be used for more enhanced and effective direct and indirect discrimination prevention dataset along with use of other differential privacy methods.

IX REFERENCES

[1] Jiuyong Lia, Hong Shenb, Rodney Topor , "Mining the optimal class association rule set" Received 2 April 2001 accepted 22 November 2001.

[2] Xiaoxin Yin Jiawei Han , "CPAR: Classification based on Predictive Association Rules " University of Illinois at Urbana-Champaign {xyin1, hanj}@cs.uiuc.edu.

[3] Asmita Kashid, Vrushali Kulkarni and Ruhi Patankar, " Discrimination Prevention using Privacy Preserving Techniques", International Journal of Computer Applications (0975 – 8887) Volume 120 – No.1, June 2015.

[4] Kamal D. Kotapalle and Shyam Gupta, "Discrimination Prevention and Privacy Preservation in Data Mining", www.ijird.com July, 2014 Vol 3 Issue 7 INTERNATIONAL JOURNAL.

[5] Rakesh Agrawal Tomasz Imielinski Arun Swami , "Mining Association Rules between Sets of Items in Large Databases" , IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120 , 1993 ACM SIGMOD Conference Washington DC, USA, May 1993.

[6] Bing Liu Wynne Hsu Yiming Ma , "Integrating Classification and Association Rule Mining", Department of Information Systems and Computer Science National University of Singapore ,Lower Kent Ridge Road, Singapore 119260.

[7] Sergey Brin Rajeev Motwani Craig Silverstein "Beyond Market Baskets: Generalizing Association Rules to Correlations", Department of Computer Science Stanford University Stanford,CA 94305.

[8] K. Ali, S. Manganaris, R. Srikant, Partial classification using association rules, in: D. Heckerman, H. Mannila, D. Pregibon, R. Uthurusamy (Eds.), Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), AAAI Press, Menlo Park, CA, 1997, p. 115.

[9] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, Fast discovery of association rules, in: U. Fayyad (Ed.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, 1996.

[10] M. Houtsma, A. Swami, Set-oriented mining of association rules in relational databases, 11th International Conference Data engineering, 1995.

[11] J.S. Park, M. Chen, P.S. Yu, An effective hash based algorithm for mining association rules, ACM SIGMOD International Conference Management of Data, May, 1995.

[12] J. R. Quinlan , "Improved Use of Continuous Attributes in C4.5", Basser Department of Computer Science, University of Sydney, Sydney Australia 2006 , Journal of Articial Intelligence Research 4 (1996) 77-90 Submitted 10/95; published 3/96

[13] P. Clark, R. Boswell, Rule induction with CN2: Some recent improvements, in: Y. Kodratoff (Ed.), Machine Learning—EWSL- 91, 1991.

[14]Fabrice Muhlenbachand Ricco Rakotomalala, "Discretization of Continuous Attributes " , Université Jean Monnet – Saint-Etienne, France.

[15]Alexandre Evfimievski,"Randomization in Privacy Preserving Data Mining", Cornell University,Ithaca, NY 14853, USA Volume 4, Issue 2.

[16] S. Hajian and J. Domingo, "A Methodology for Direct and Indirect Discrimination prevention in data mining." IEEE transaction on knowledge and data engineering, VOL. 25, NO. 7, pp. 1445-1459, JULY 2013.

Author's Profile

Mr. Krishna Kumar Tripathi pursed Master of Engineering from Mumbai University in year 2013. He is currently pursuing Ph.D, and currently working as Assistant Professor in Department of Computer Enginnering, He has 16 years of teaching experience.

Dr. Narendra S. Chaudhari, is currently working as professor and director of Visvesvaraya National Institute of Technology Nagpur in Department of Computer Engineering, India. He has 30 years of experience. Dr. Narendra S. Chaudhari has shouldered many senior level administrative positions in universities in India as well as abroad. He has many publications in top quality international conferences and journals.