International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.101.16

Data Restructuring with Feature Set Partitioning for Privacy Preserving using K-anonymity

Md. Riyazuddin[1], Dr. V.V.S.S.S. Balaram[2]
[1]Asst. Prof., MJCET, Hyderabad.,
[2]Professor, SNIST, Hyderabad.

Abstract : Most of the data available for knowledge discovery and information retrieval are prone to identity and privacy disclosure. The major act to disclose the identity is through exploring the pattern of attributes involved in data formation. The existing benchmarking models are anonym zing the data either by generalizing, deleting the sensitive attributes or adding noise to the data. Either of these approaches are not guaranteed in optimality and accuracy in results that obtained from the mining models applied on that dataset. The deviation in results often causes falsified decision making, which is unconditionally not acceptable in few domains like health mining and real time environment. In order to fill the gap, here we proposed a hybridization of feature set partitioning and data restructuring to achieve the pattern anonymization. The model is particularly aimed to restructure the data for supervised learning. To the best of our knowledge, pattern anonymization is first of its kind that attempted to anonymize the patterns rather individual attributes. The experiment results also indicating the scope of robustness and scalability of the supervised learning on restructured data.

Keywords: Privacy preserving, Data Mining, Data Anonymization, Pattern Anonymization, Identity Disclosure, Data Restructuring, Feature set Partitioning, K-anonymity, l-diversity.

I. INTRODUCTION

The exploration of eligible data patterns [1] is a key process of information retrieval and Knowledge discovery. Data mining is one of the significant concept, which is generally used for information retrieval and knowledge discovery by identifying the eligible patterns from the given data [2], [3], [4]. The considerable consequence of this data mining algorithms and models is identity and personal information discloser.

Above 90% of the participants of a survey [5] aware that data related to any individual can be shared or sell by a company only under the permission given that individual, henceforth the sharing of personal information by an individual to an eligible organization is hassle free. The data mining applications often applies on these personal and sensitive data leads to violate individual privacy. Henceforth, often the implementations of these mining models are restricted [6], [7], [8].

The objective of this study is related to securing the privacy of the data from leakages possible in data mining. The organizations should protect the privacy of the customer's personal information, which could be revealed unofficially due to patterns discovered by data-mining activities [9].

It has been learnt that removal of sensitive identities related to personal information is not adequate solution for privacy protection [10].The data publishing for statistical analysis is another domain, which is sensitive to this privacy leakage. The considerable study was done and landed with solutions like restricting queries and perturbation the data to protect privacy in data publishing [11]. But none of these methods are compatible to preserve privacy in data mining.

The predictive mining strategies such as classification cannot be possible on query restricted and perturbed dataset. Since the predictive mining models require to explore the possible associations between attributes. Henceforth anonymizing the data is considered to be the best approach to prevent privacy leakages in predictive mining models. Many of such solutions [12],[13],[14],[15],[16] can be found in the literature of past decade. The most successful privacy protection strategy was k-anonymity [17], [18] that anonymizes the record of each individual, such that it can't be distinguished from minimum K other individuals. This is done mostly by generalizing the sensitive attributes of the records or eliminating these sensitive attributes from the given dataset or including noise such that no individual personal information is leaked.

The similar objective has been considered in this research article. The aim is to explore the constraints of the existing models and defining a data anonymization approach towards privacy preserving data mining.

The rest of the paper is organized as follows. The section 2 contains the exploration of the strengths and constraints of the existing benchmarking models and that followed by section 3 and 4, which contains detailed projection of the proposed model. The section 5 presenting the experimental setup and performance analysis that followed by the section 6, which concludes the proposal.

II. RELATED WORK

K-anonymity is optimized in [19], which is done by performing a search to identify the attributes possible to allow the privacy leakage. A multi-objective method for hiding sensitive association rules is devised in [20]. This model is an evolutionary strategy since it is using GA to identify the sensitive rules. This model is optimal to preserve privacy and deliver extremely significant rules. The main constraints of these models [19][20] are the need of prior knowledge of the sensitive attributes of the dataset and computational complexity of search is nonlinear.

The concept of feature set partitioning introduced in [21] is aimed to magnify the scalability of the supervised learning. The feature set partitioning decomposes the actual feature set into multiple subsets and further builds each subset level classifier. Further ensembles all these classifiers to recognize the class of the target record [22]. Generalizing the feature during the feature selection is the main objective of feature set partitioning strategy. The classifier will be constructed by using the representative features. The empirical study of the model evinced that the optimality is proportionate to the formation of minimal number of subsets from maximal number of features, which is also a significant constraint for datasets with sparse and divergent feature set. The privacy leakage is another biggest constraint of this supervised learning by feature set partitioning.

In order to overcome the constraint of privacy leakage observed in classification strategy called feature set partitioning [21], Matatov et al., [23] extended the model devised in [21]. The model proposed in [23] is achieving optimal K-anonymity for each feature subset to prevent identity leakage of the dataset. This model is also an evolutionary strategy as it is using genetic algorithm to identify the optimal feature subsets. The empirical study evincing that the model is optimal to achieve K-anonymization for defined feature subsets. The considerable limits of the model are nonlinear computational complexity in optimal feature subset discovery and requirement of prior domain knowledge to identify the feature's sensitivity towards privacy leakage, which is an essential factor to define cost function of the genetic algorithm. The anonymization is relevant and specific to the ensemble classification by partitioned feature set proposed in [21]

The existing benchmarking models that aimed to privacy preserving datamining are anonymizing data by generalizing, obliterating the quasi fields of the records in given dataset or restructuring the dataset by adding noise. The process of generalizing or obliterating the field values may achieve optimal anonymity but causes severe violation in mining results. On other method of adding noise also causes considerable deviation in mining results, which is due to the noise included. Another specific hurdle in any of the existing models is the compulsion of the prior knowledge about the quasi attributes. Hence the any of the benchmarking models minimizes their negative impact on mining results, if and only if the anonymization done in the context of the dataset under the close monitoring of domain experts and specific to particular mining algorithm. In most of the cases prior knowledge of the data and close monitoring of the domain experts is big constraint. Hence it is obvious to conclude the need of an optimal strategy for privacy preserving data mining that works without prior knowledge of the data given and close monitoring of the domain experts. On other dimension the anonymization to pre-serve the privacy should not violate the originality of the mining results.

In order to this, here in this article we proposed a hybrid approach that combines the feature set partitioning and anonymizing through dataset restructuring by including trivial records. The proposed model is anonymizing the patterns observed in given dataset, since the feature patterns are the primary factors those leads to identity leakage. Also the other unique feature of the proposal is that it compatible to any of the supervised learning based mining algorithm. To the best of our knowledge, model defined here is first of its kind that retains structure of the actual records while adding the trivial records to achieve pattern disclosure.

III. PATTERN ANONYMIZATION BY FEATURE SET PARTITIONING AND DATASET RESTRUCTURING

3.1 Notations used in model exploration:

- Dataset: A set of records and each record contains values for fixed number and order of attributes associating to a class label.
- Class Label: An attribute representing the state of the record
- Attribute set: A set of attributes representing few or all attributes those labels the different fields of the records.
- Value set: The set of values in a record representing an attribute set is known as Value set

 Trivial Record: A record that contains trivial values for one or more attributes.
 Trivial Value Set: A set of values in a trivial record representing an attribute set such that one or more values are trivial.

3.2 Proposed Work:
The optimal pattern-preserving k-anonymization problem can be formulated
as follows:
Definition 1 (optimal P2kA problem).
Given a sequence database D, and a positive integer k, find a database D' such that
1. D' is k-anonymous, i.e.:

$$\sum_{T \in \mathcal{S}(\mathcal{D}')} \delta[supp_{\mathcal{D}}(T) < k] \cdot supp_{\mathcal{D}'}(T) = 0$$

2. the collection of all k-frequent pattern in D is preserved, i.e.:

$$\mathcal{S}(\mathcal{D}', k) = \mathcal{S}(\mathcal{D}, k)$$
$$\forall T \in \mathcal{S}(\mathcal{D}', k)\ supp_{\mathcal{D}'}(T) = supp_{\mathcal{D}}(T)$$

In this paper we present an algorithm which assures that
(i) D' is k-anonymous and
(ii) (ii) S (D, k) and S (D', k) are "similar". In particular the second condition of Definition 1 becomes:

$$\mathcal{S}(\mathcal{D}', k) \subseteq \mathcal{S}(\mathcal{D}, k)$$
$$\forall T \in \mathcal{S}(\mathcal{D}', k)\ supp_{\mathcal{D}'}(T) \simeq supp_{\mathcal{D}}(T)$$

Algorithm 1: BF-P2Ka (D, K)
Input : A sequence database D, a min. support threshold k
Output: A k-anonymous sequence database D'
PT = PrefixTreeConstruction (D);
PT' = PT Anonymization (PT, k)
D' = SequenceGeneration (PT');
Return D'
Algorithm 2 : PTAnonymization (PT, k)
Input : A prefix tree PT, a min. support threshold k
Output : A k-anonymous prefix tree PT'
L_{cut}=0;
For each n in Root (PT) children do
$L_{cut\ t}$= L_{cut} U TreePruning(n,PT,k);
End
PT' = TreeReconstruction (PT, L_{cut});
Return PT'

3.3 Feature set Partitioning
Let consider all the attributes of the given dataset as a set A . Further all possible subsets from A of different sizes will created as a set S .
Similarly create a set R that contains the subsets formed from each record r of the given dataset D , which is done based on the subsets defined from A . A subset rs_i contains the values from r of the attributes found in subset{s_i s_i S} .
Find coverage of each subset {r_j s_i r_j D s_i S} , which represents the number of records contains r_j s_i
Prune the subsets from R under bide rule [citation required] as follows:
Let a set r_j s_i with coverage k and a set r_p s_q with coverage k ' , if r_j s_i r_p s_q and
k k ' then r_j s_i can be discarded.

3.4 Dataset Restructuring
Let K be the optimal number of records representing each feature set partition to achieve K-anonymity.
If any of the value set r_j s_i from R , which is representing values for attribute set s_i in record r_j is with coverage k less than K then, K k number of trivial records will be formed such that each trivial record with trivial value set tr_j s_i for attribute set s_i replaced by respective value set r_j s_i . This is done by updating existing trivial records if any otherwise creates new trivial records.

Further each trivial record will be associated to a class label randomly from all pos-sible class labels, which is done in order to achieve maximum possible diversity in class label representation.

IV. PROCESS MODEL OF THE FEATURE SET PARTITIONING AND DATA RESTRUCTURING

For set of size n , possible number of subsets are 2^n □1, which excludes the empty set [citation required]. Hence forth,2^{n} □1 attribute sets S □ {s , s , s ,.....s from |S| 1 23 |} find the set of attribute labels A of size n , which excludes class label attribute.

Order S by the size and order of attributes
Let R be an empty set
For each record {r_i □ r_i □ D□i □ 1,2,...,| D|} that represents the values for all attributes of the set A Begin
Prepare 2^{n} □1 value sets r S □ {r s , r s ,..., r s r } from values of record that exi i 1 i i 2 i | S | cludes the class label value. Order r_i S in the order of S
R □ r_i S //move r_i S to R
End

4.1 Further find all possible subsets from all entries of the R as follows:
Let $\overline{R}$ be the empty set
For each {r_i S □ r_i S □ R□ i □ 1,2,3.... | D |} Begin
R □ □R r_i S□
End
Let C be the set of all possible unique class labels observed for records in dataset D

4.2 Further find the coverage of the each entry of R as follows
k $\overline{R}$ □ □ // is an empty map
l R □ □ // is an empty map
For each □rs_i □rs_i □ R□i □ 1, 2,|..,| R □ Begin
For each □c □c □ C□ Begin
// C is a set of class label {c_i □c_i □ C □i □ 1, 2,...,| C |}
l □ 0 k □ 0
For each {r_j □r_j □ D□ j □ 1,2,...,| D |} Begin
If □rs_i □ r_j □ then Begin
k □ k □1
If □r_j (cl) □ c□ then begin
// cl is the class label of the record r_j
l □ l □1
End
End
End
k R{rs_i } □ k
l R{rs_i (c)} □ l
End
End
Sort R in ascending order of the subset size and order of attributes

4.3 Further Prune l R and k R as follows

For each □rs$_i$ □rs$_i$ □ R□i □ 1, 2,..., R □1□Begin
For each □rs$_j$ □rs $_j$ □ R□j □ i □ 1,i □ 2,..., R □ Begin
If □□k R{rs$_i$ } □ k R{rs $_j$ }□ & &□rs$_i$ □ rs $_j$ □□ then Begin
// delete key rs$_i$ and respective value from k R
k R □ k R \{rs$_i$ }
End
For each □c$_m$ □c $_m$ □ C □m □ 1, 2,...,| C |□ Begin
If □□l R{rs$_i$ (c$_m$)} □ l R{rs $_j$ (c$_m$)}□& &□rs$_i$ □ rs $_j$ □□ then Begin // delete key rs$_i$ (c$_m$) and respective value l R \ rs$_i$ (c$_m$)
End
End
End
End

Let {keys}$_k$ be the set of all keys exists in k R
Let {keys}$_l$ be the set of keys exists in l R

//Anonymity
Let K be the expected number of records representing a given feature partition fp for optimal anonymity
//Diversity
The number of records representing a given feature partition fp associable to class labels {C } □ c for optimal anonymity.

4.4 Restructuring Data for K-Anonymity of the feature sets is follows:

// Sort {keys}$_k$ in ascending order of the size of all subsets exists as keys Let D □ □ //The new empty dataset.
//A record □tr □tr □ D□ that contains trivial value for at least one attribute {a □a □ A}

is called trivial record tr .

// the number of attributes contains trivial values in tr indicates |tr|
□ {key}$_k$ □i □ 1, 2,...,| {keys}$_k$
For each {rs$_i$ □rs$_i$ |} Begin
{rs$_i$ } □ K
If □k R□ then Begin
p □ K □
k R{rs$_i$ }

D □ □tr
If □D □ □ □ then□ tr |□ A|□
// tr is a record with trivial values for all attributes of A
idx □ 0

□ □ tr □tr □ D□&□
□ & □
□ □{keys}$_k$ □ &
□ trs$_i$ &□
For each trivial record □ □ Begin
□ □
□ □tr □| trs$_i$ | □ & &□
□ □
□ □ idx □ p□ □

// trs_i is the values set for subset $\{s_i \in s_i \in S\}$ in tr ,
$\exists\{keys\}_k \in$ indicates that all values in
// $\exists trs_i$ trs_i are trivial and

$|\exists| trs_i |\in$ indicates that number of attributes
// $\in$ tr in tr with trivial values must be

greater than the size of the trs_i

$\{trs_i \exists trs_i \in tr\} \rightarrow rs_i$ // replacing trivial values of the set trs_i by the values of the set

rs_i

idx $\leftarrow$ idx $\exists$1 End
If $\exists$idx $\in$ p$\exists$ then for each $\{q \exists q \in idx , idx \in 1,idx \in 2,..., p\}$

$\bar{D} \in \exists tr \| \in \| tr_{|A|}$ is a record with trivial values for all attributes
$\in$ A $\exists$// tr of A
$\{trs_i \exists trs_i \in tr\} \rightarrow rs_i$ // replacing trivial value set trs_i of record tr by the value set

rs_i

End

End

End

4.5 Add label to all trivial records in $\bar{D}$ is as follows:

For each $\exists tr \exists tr \in \bar{D} \in$ Begin

// assigning a label selected randomly from class label set, this is done to achieve maximal diversity of associability between attribute values and class labels
tr (cl) $\leftarrow$ rand ({C}) End

D $\leftarrow \bar{D}$ // adding all the records from D to restructured dataset D

V. EXPERIMENTAL SETUP AND RESULTS ANALYSIS

The experiments were conducted to assess the compatibility of the restructured data set towards supervised learning. The impact of K-anonymity with maximal possible diversity is a proven strategy towards privacy preserving [24]. Hence the experiments con-ducted here were not aimed to explore the optimality of the K-anonymity and maximal diversity.

The accuracy, robustness and scalability of the results obtained from restructured dataset are assessed through statistical metrics [25] called precision, sensitivity and ac-curacy, which are estimated by using the count of truly classified and count of falsely classified.

Since the assessment metrics called computational and resource complexity also included in performance analysis, a computer with i5 processor, 4GB ram and Nvidia 4GB graphics card is used. The implementation was done in CUDA [26]. Statistical metrics analysis was done using explorative language R [27]. The input and obtained results were explored in table 1.

5.1 The dataset

The objective of the proposed model (Pattern Anonymization Approach) is to perform the optimal supervised learning on restructured dataset that protects from pattern disclosure. To assess the scalability and supervised learning accuracy, we adopt the heart disease dataset [28]. We initially classified the dataset by classification tool J48 [29] and obtained prior knowledge of the possible groups of records.

5.2 Performance Analysis

The classes predicted by the proposed pattern anonymization approach (PAA) were assessed, by comparing the classification of the original records under restricted dataset, which includes trivial records also. The Metric values indicating that classification of original records after restructuring the dataset is significant (precision is 0.99339934 that indicates the truly classified records ratio). The sensitivity of true and trivial record classification is also considerably high (sensitivity is 1 that indicates no trivial records included into actual groups of the original records). The overall classification optimality is observed as best, since, almost the 100% of the records grouped into relevant labels under the given restructured dataset and experimental setup (accuracy is 0.995).

No of features	18
No of records in original dataset	404
No of records in restructured dataset	616
No of groups formed from original records	18
No of groups formed from restructured dataset (original and trivial records)	25
No of groups after pruning the trivial records	14
Truly classified no of records	402
Falsely classified records	2
Precision	0.99504950
Sensitivity	1
Accuracy	0.995

Table 1. Particulars of the input dataset and results obtained

The computational complexity and resource cost is also assessed, which is done under divergent count of records given as input. The time complexity observed to be linear for given divergent count of records as input (see fig 1). The memory usage of Dataset restructuring with Feature set partitioning also being noticed as linear for given diver-gent count of input records (see fig 2).

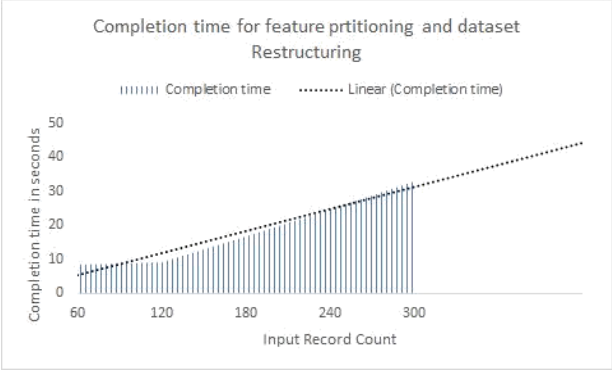

Fig. 1. Figure 1: Feature set Partitioning and dataset restructuring completion time observed for divergent count of input records

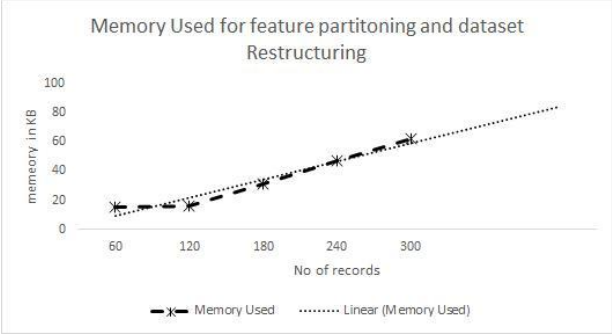

Fig. 2. Memory used for Feature set Partitioning and Dataset Restructuring of divergent count of input records

VI. CONCLUSION

Privacy preserving in supervised learning is prime objective of the model proposed here in this paper. In this context many of existing models succeeded to prevent privacy disclosure under certain constraints such as prior knowledge of the data to identify the sensitive attributes, compromising at optimality of the mining results due to sensitive attribute eradicating and generalizing or nonlinear complexity observed in the process of sensitive feature identification. The proposed model is hybridizing the data restructuring with feature set portioning to achieve privacy preserving for any of existing supervised learning based mining model. The best part of this model is that no prior knowledge of the data is required to anonymize and computational complexity is observed was linear. The important factor to adopt this model is that it can't violate the mining results, which is a biggest constraint of the existing models. The motivation gained from this model drives our future research to minimize the computational complexity to much minimal that compared to present model. In other direction of future research the similar anonymization strategy can be devised for rule mining and unsupervised learning strategies.

VII REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, in: Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 1–31.
[2] H. Chen, Intelligence and security informatics: information systems perspective, Decision Support Systems 41 (3) (2006) 555–559.
[3] D. Martens, L. Bruynseels, B. Baesens, M. Willekens, J. Vanthienen, Predicting going con-cern opinion with data mining, Decision Support Systems 45 (4) (2008) 765–777
[4] T.S. Raghu, H. Chen, Cyberinfrastructure forhomeland security: advances in information sharing, data mining, and collaboration systems, Decision Support Systems 43 (4) (2007) 1321–1323.
[5] S. Greengard, Privacy: entitlement or illusion? Personnel Journal 75 (5) (1996) 74–88.
[6] M. Kantarcioglu, J. Jin, C. Clifton, When do data mining results violate privacy?, in: Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, 2004, pp. 599–604.
[7] C. Clifton, M. Kantarcioglu, J. Vaidya, Defining privacy for data mining, in: H. Kargupta et al. (Eds.), Proc. of the National Science Foundation Workshop on Next Generation Data Mining, Baltimore, Maryland, 2002, pp. 126–133.
[8] M. Feingold, M. Jeffords, M. Leahy, Data Mining Moratorium Act of 2003, US Senate Bill (proposed), 2003.
[9] L. Cao, C. Zhang, Domain-driven, actionable knowledge discovery, Intelligent Systems 22 (4) (2007) 78–88.
[10] Zhu, Dan, Xiao-Bai Li, and Shuning Wu. "Identity disclosure protection: A data reconstruc-tion approach for privacy-preserving data mining." Decision Support Systems 48.1 (2009): 133-140.

[11] N.R. Adam, J.C. Wortmann, Security-control methods for statistical databases: a compara-tive study, ACM Computing Surveys 21 (4) (1989) 515–556.
[12] A. Amiri, Dare to share: protecting sensitive knowledge with data sanitization, Decision Support Systems 43 (1) (2007) 181–191
[13] D.S. Chowdhury, G.T. Duncan, R. Krishnan, S.F. Roehrig, S. Mukherjee, Disclosure detec-tion in multivariate categorical databases: auditing confidentiality protection through two new matrix operators, Management Science 45 (12) (1999) 1710–1723
[14] R. Garfinkel, R. Gopal, P. Goes, Privacy protection of binary confidential data against de-terministic, stochastic, and insider threat, Management Science 48 (6) (2002) 749–764.
[15] S. Menon, S. Sarkar, Minimizing information loss and preserving privacy, Management Sci-ence 53 (1) (2007) 102–116.
[16] S. Menon, S. Sarkar, S. Mukherjee, Maximizing accuracy of shared databases when con-cealing sensitive patterns, Information Systems Research 16 (3) (2005) 256–270.
[17] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.
[18] L. Sweeney, k-Anonymity: a model for protecting privacy, International Journal on Uncer-tainty, Fuzziness and Knowledge-based Systems 10 (5) (2002) 557–570.
[19] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k-Anonymization. Proceedings of the ICDE Conference, pp. 217–228, 2005.
[20] Dehkordi, M. N., Badie, K., & Zadeh, A. K. (2009). A novel method for privacy preserving in association rule mining based on genetic algorithms. Journal of software, 4(6), 555-562.
[21] L. Rokach, O. Maimon, Theory and application of feature decomposition, in: Proc. of the First IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, 2001, pp. 473–480.
[22] E. Menahem, L. Rokach, Y. Elovici, Troika – an improved stacking schema for classifica-tion tasks, Information Sciences 179 (24) (2009) 4097–4122.
[23] Matatov, N., Rokach, L., & Maimon, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. Information Sciences, 180(14), 2696-2720.
[24] Chen, Rui, et al. "Privacy-preserving trajectory data publishing by local suppression." Infor-mation Sciences 231 (2013): 83-97.
[25] Powers, D. M. (2006). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 23rd International conference on machine learning. Pitsburg.
[26] Nvidia. (2008). C. U. D. A. Programming guide.
[27] Ihaka, R. &. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statistics, 299-314.
[28] https://archive.ics.uci.edu/ml/machine-learning-
[29] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classifica-tion algorithm for data classification." International Journal of Computer Science and Ap-plications 6.2 (2013): 256-261.
[30] Pattern Anonymization: Hybridizing Data restructure with Feature Set Partitioning for privacy preserving in Supervised Learning. By MD. Riyazuddin and Dr. V.V.S.S.S. Balaram, ICCII 2016, SPRINGER AISC series, Singapore 2016, pp 603-614.
[31] Pattern Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining. By Ruggero G. Pensa1, Anna Monreale, Fabio Pinelli1 and Dino Pedreschi.