PRESERVING PRIVACY USING DATA ANONYMIZATION FOR KNOWLEDGE DISCOVERY

[1]Jalpa Shah, [2]Jayshree Upadhyay
[1]Assistant Professor, [2]Assistant Professor
[1]Computer Engineering,
[1]Aditya Silver Oak Institute of Technology, Ahmedabad, India

Abstract—In this information age, data and knowledge extracted by data mining techniques represent a key asset driving research, innovation, and policy-making activities. Many agencies and organizations willing to release the data they collected to other parties, for purposes such as research and the formulation of public policies. The success of data mining relies on the availability of high quality data. To ensure quality data mining, effective information sharing between organizations becomes a vital requirement in today's society. Since data mining often involves data that contains personally identifiable information and therefore releasing such data may result in privacy breaches; this is the case for the examples of micro data, e.g., census data and medical data. Privacy preserving data publishing (PPDP) is a study of eliminating privacy threats like linking attack while, at the same time, preserving useful information in the released data for data mining. Privacy Preserving Data Mining (PPDM) field of research studies how knowledge or patterns can be extracted from large data stores while maintaining commercial or legislative privacy constraints. Quite often, these constraints pertain to individuals represented in the data stores. This work is about proposing a method which extends the process of anonymization to achieve new knowledge through data mining while protecting individuals' privacy.

***Index Terms*— Anonymization, Privacy preserving Data Mining**

I. INTRODUCTION

There are so many organizations who publish their data in various forms. These forms contain various information. Information can be helpful for someone and at the same time can be useless for another one. Some information may be important for business point of view, industrial point of view that depends on person to person. So which information is sensitive i.e. we do not want to disclose it for general people and which information can be published. So caring of these issues, organization needs to publish their information. As for example in a hospital system a lot of patient comes for their treatment in respective departments. Hospital need to maintain their records and make a file for that which contains patient information. They want to publish reports such that information remains practically useful and the important thing is that identity of an individual cannot be determined. So publishing of data is main concern here. Organization needs to publish microdata. Microdata e.g. Medical data, voter registration and census data for research and other purposes. These data are stored in a table. Each record corresponds to one individual [1]. Microdata is a valuable source of information for the allocation of public funds, medical research, and trend analysis. However, if individuals can be uniquely identified in the microdata then their private information (such as their medical condition) would be disclosed, and this is unacceptable. Each record has number of attributes, which can be divided into three categories. (1) Explicit identifiers attributes that clearly identify an individual. E.g. - social security number. (2) Quasi-identifiers attributes whose value when taken together can identify an individual. e.g. Zip-code, birth date and gender. (3) Attributes those are sensitive such as disease and salary. It is necessary to protect sensitive information of individuals from being disclosed. There are two types of information disclosure identity disclosure and attribute disclosure [9]. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. If there is only one female black dentist is in area and sequence queries reveal that she is in database then identification occurs. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, to limit the disclosure risk to an acceptable level while maximizing the benefit. This can be done by anonymizing the data before release. By knowing the quasi identifiers can lead to know the sensitive information. This can be done by knowing the individual personally or other publicly available database [2].

II RELATED WORK

Databases today can range in size into the terabytes of data. Within these masses of data lies hidden information of strategic importance. The newest answer is data mining, which is being used both to increase revenues and to reduce costs. The potential

returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud [3].

Data mining (also known as KDD – Knowledge Discovery in Database) is a process that uses a variety of data analysis tools to discover patterns (finding interesting information) and relationships in data that may be used to make valid predictions.

PPDM aims at providing a trade-off between sharing information for data mining analysis, on the one side, and protecting information to preserve the privacy of the involved parties on the other side. PPDM approaches protect data by modifying them to mask or erase the original sensitive data that should not be revealed.
PPDM approaches are based on principle- loss of privacy, measuring the capacity of estimating the original data from the modified data, and loss of information, measuring the loss of accuracy in the data. The main goal of these approaches is therefore to provide a trade-off between privacy and accuracy.
In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002). [14]
Individual privacy preservation: The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual [4].

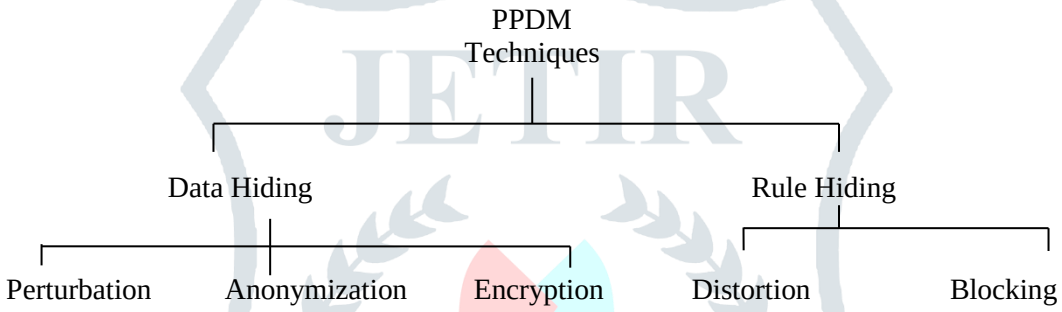

Figure 2.1: PPDM Techniques

Many researchers have found several approaches for data preservation for data publishing. So several methods such K-anonymity, L-diversity, T-closeness and others are come into existence to maintain privacy in data publishing. In this paper we discussed pros and cons of all these techniques [5].

We have introduced the problem of privacy-preserving data stream mining and discussed the broad areas of research in the field. The broad areas of privacy are as follows:

- Since the perturbed data may often be used for mining and management purposes, its utility needs to be preserved. Therefore, the data mining and privacy transformation techniques need to be designed effectively, so to preserve the utility of the results.
- Running time is still an open issue to many of the algorithms. There are so many anonymization algorithms but each have different time for execution.
- How to identify a proper quasi-identifier is a hard problem as it depends on what the external table looks like. Also it is hard to predict what external tables will be used to inference the sensitive information.
- The cost of K-Anonymous solution to a database is the number of *'s introduced. Hence to find a k-anonymity solution with suppressing fewest cells is very critical.
- How to generate a table with less distortion while performing fast is still open issue.
- Extending ideas for handling multiple sensitive attribute, and developing methods for continuous sensitive attributes
- Performance improvement in proposed algorithm is still an open issue.
- Generalized algorithm for both categorical and numerical values poses more challenge
- The curse of dimensionality becomes especially important when adversaries may have considerable background information

III. PROBLEM STATEMENT

Data Privacy and Data accuracy can be compared with see-saw, i.e. if you increase privacy there less data accuracy and vice versa. As we have seen the entire PPDM algorithm, none of these provide individual tuple based privacy gain.
For given any microdata dataset T. Find the Quasi-identifier from it. Apply anonymization on it with minimum perturbation to increase result accuracy. If information loss is minimum then the mining result on that is might be accurate like original result. So perform mining on original dataset T, apply anonymization technique to convert it into T', then again apply data mining on T'. After outcome of that compares the result of both T and T'. You will find some information loss and that should be minimum.
We can compare the classification characteristics in terms of less information loss and more privacy gain. So get better accuracy.

IV. PROPOSED SOLUTION

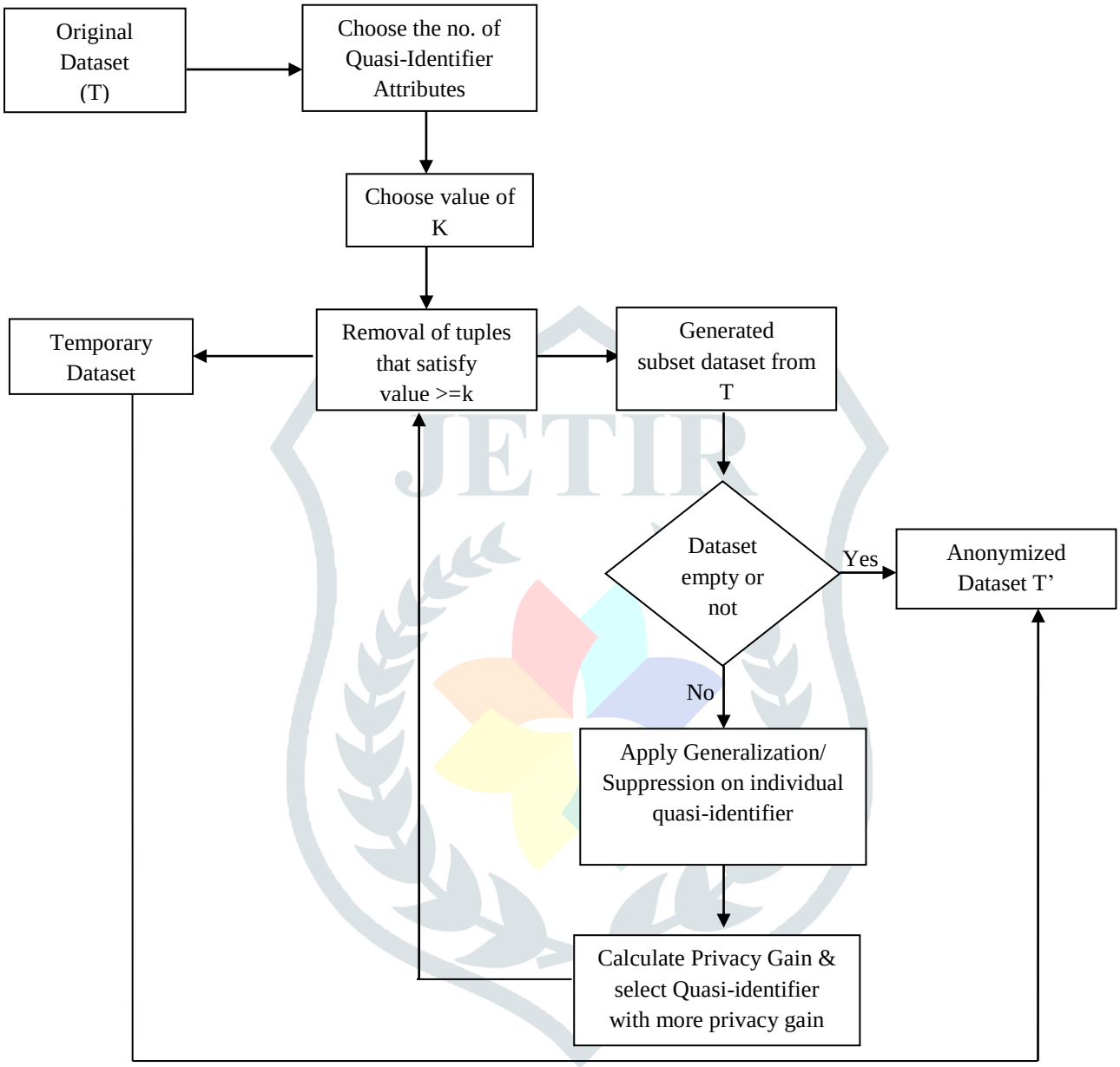

Figure 4.1: Proposed architecture

Proposed Algorithm:
Input: Original Dataset T
Output: Anonymized Dataset T'
Step:

1. Take dataset T
2. Take the value of Quasi-identifier from user
3. Take the value of ' k ' from user
4. Remove (write in output Dataset T') all tuples having value >= k
   Also make subset of all tuples having value < k (after removing/ write to output dataset)
5. While D is not empty
   a. Make subset (all tuples having value < k) S and apply processing on this subset
   b. Take any attribute or set of attribute for anonymization to make similar tuples
   c. In processing, you have to make a group of similar tuples, if not possible then anonymize it so, if tuple have age attribute, generalize it, means convert age 4, 3 into group like [0-5]
   d. Now, again check the each tuples of subset and match it with k, if it satisfies the value of, go to step-3
6. Again store the tuple having value <k to make new subset S'

7. Apply more anonymization (take another attribute like gender, then another attribute like pincode, then set of two like age and Salary, or age and pin code or Salary and pin code) And see the frequency of tuple until Dataset T is empty or no subset is left.
8. If still any tuple in subset is left then do full domain generalization and put it into output Dataset T'
(Anonymize only what is required, not more than that)

V. IMPLEMENTATION

We have conducted experiments to evaluate the performance of data Anonymization method. We choose two databases. We use Waikato Environment for Knowledge Analysis (WEKA) tool to test the accuracy of *Naive Bayes algorithm*. The data Anonymization algorithm implemented by a separate Java program. Here in experimental setup we are using the tool –WEKA for experimental analysis purpose [9].

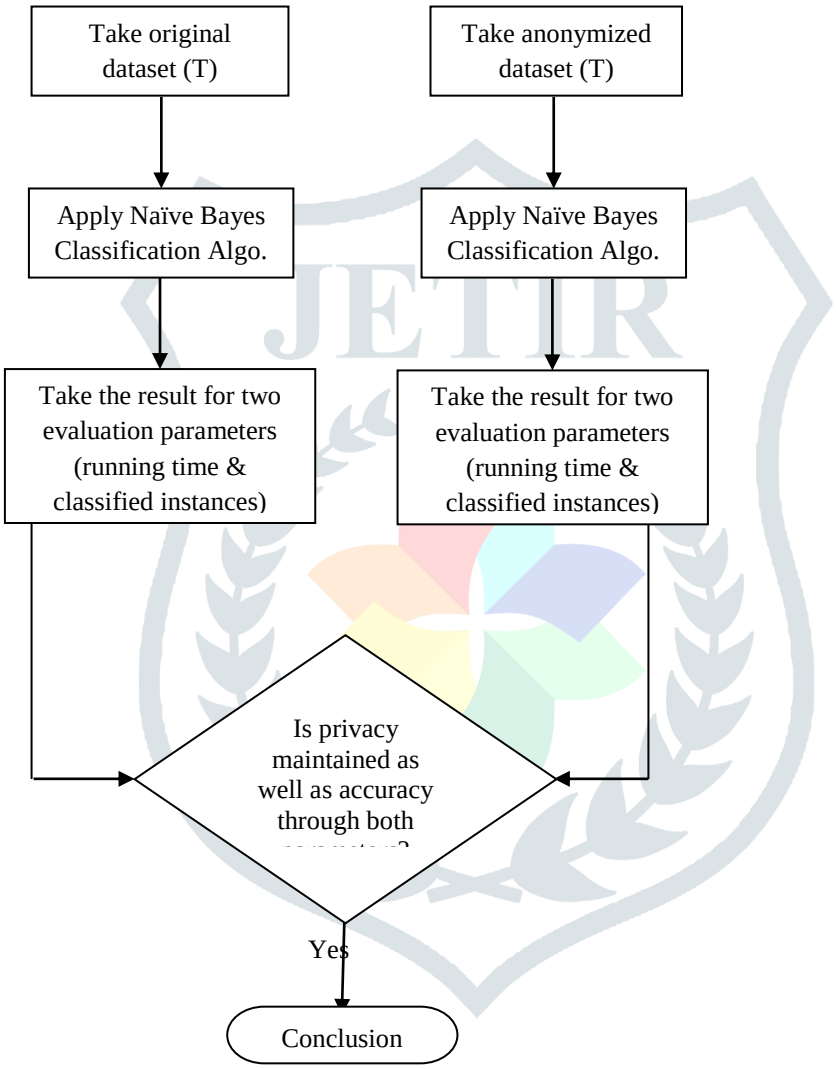

Figure 5.1: Methodology of Experiment

After doing the implementation of proposed algorithm, the comparison of results for time taken to build model and correctly classified instances is to be analyzed.

Table: 5.1 Experiment result (q=2 with lower value of k)

Dataset	Analysis	Original Dataset	Anonymized Dataset		
			k=2	k=3	k=4
Adult Dataset	Time taken to build a model	0.23	0.17	0.17	0.17
	Correctly classified Instances	100	97.3087	97.1934	97.0985
Bank Marketing Dataset	Time taken to build a model	0.39	0.27	0.25	0.25
	Correctly classified Instances	100	96.33947	96.3284	96.2797

Table: 5.2 Experiment result (q=2 with higher value of k)

Dataset	Analysis	Original Dataset	Anonymized Dataset		
			k=5	k=6	k=7
Bank Marketing Dataset	Time taken to build a model	0.39	0.25	0.25	0.23
	Correctly classified Instances	100	96.2908	96.2819	96.3438

Table: 5.3 Experiment result (q=3)

Dataset	Analysis	Original Dataset	Anonymized Dataset		
			k=2	k=3	k=4
Adult Dataset	Time taken to build a model	0.23	0.15	0.16	0.15
	Correctly classified Instances	100	96.4466	96.3668	96.4589
Bank Marketing Dataset	Time taken to build a model	0.39	0.22	0.2	0.19
	Correctly classified Instances	100	93.2296	94.0325	93.6388

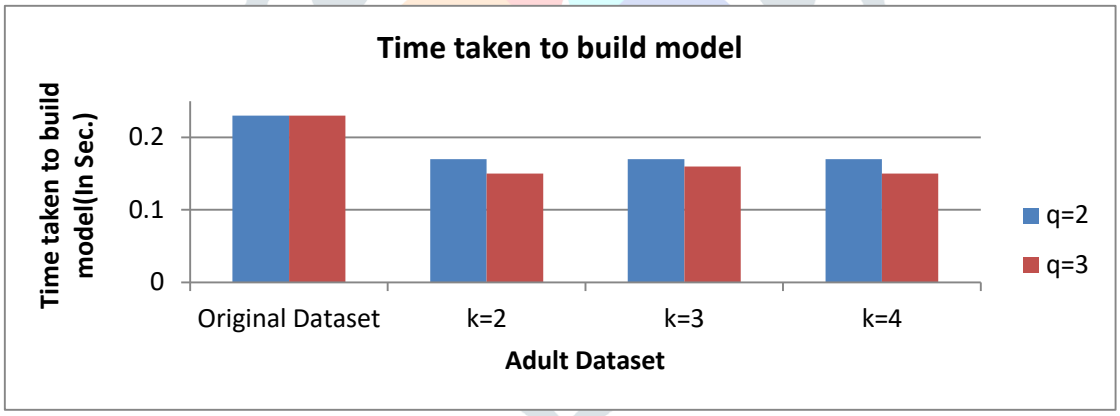

Figure: 5.2 Time taken to build the model (Adult Dataset)

As shown from the graph, that our results with different value of k and q, it is nearly equal to the original results. There are results some less than then the original, that is because of the Anonymization or due to privacy gain. So it's better to use this when you want to provide privacy to individual or on to attribute disclosure. One more thing is noticed here is the time taken to build model on anonymized dataset is less than the time taken to build original dataset [10].

CONCLUSION

After the implementation of proposed algorithm, the results are analyzed by comparing time taken to build model and correctly classified instances. We can conclude that our results with different values of k and q, it is nearly equal to the original results. So after taking the result we can conclude that time taken to build anonymized dataset is less than the original dataset, other the correctly classified instances(misclassification error) is almost negligible(3 to 5%). So we can say that while preserving privacy accuracy of dataset is maintained.

REFERENCES:

[1] D. Lambert, "Measures of Disclosure Risk and Harm," *Journal of Official Statistics,* vol. 9, pp. 313-331, 1993.
[2] Ayre, L.B.: *Data Mining For Information Professionals.* (2006)
[3] Clifton, C., Kantarcioglu, M., and Vaidya,J.: *Defining Privacy for Data Mining*. In: Next Generation Data Mining, AAAI/MIT Press. (2004)
[4] Agrawal, R., Srikant, R.: *Privacy preserving data mining.* In: Proceedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. (2000)
[5] Oliveira, Stanley R.M.: Privacy-Preserving Data Mining-*Encyclopedia of Data Warehousing and Mining*, Second Edition. IGI Global, pp. 1582-1588. (2009)
[6] Sweeney, L.: *Achieving k-anonymity privacy protection using generalization and suppression.* In: International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), pp. 571–588. (2002)
[7] Ninghui Li,tiancheng li and suresh venkatasubramanian. "Closeness: a new privacy measure for data publishing".IEEE, july 2010.
[8] G. T. Duncan and D. Lambert, "Disclosure-Limited Data Dissemination," *Journal of The American Statistical Association,* vol. 81, pp. 10-28, 1986.
[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "`l-Diversity: Privacy Beyond k-Anonymity," *Proc. Int'l Conf. Data* Engineering *(ICDE),* pp. 24, 2006.
[10] T. M. Truta and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property," *Proc. Int'l Workshop on Privacy Data Management (ICDE Workshops),* 2006.
[11] Bayardo, R. and Agrawal, R.: *Data privacy through optimal k-anonymity.* In: Proceedings of the 21st International Conference on Data Engineering (ICDE), 6(2), pp. 217-228. (2005)
[12] Wong,R., Li, J., Fu, A., and Wang, K.: *(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing*. In: Proceedings of the 12th ACM SIGKDD, pp. 754-759. (2006)
[13] LeFevre, K., DeWitt, D., Ramakrishnan, R.: *Incognito: Full Domain K-Anonymity*. In: ACM SIGMOD Conference, pp. 49-60. (2005)
[14] LeFevre, K.,DeWitt, D., and Ramakrishnan, R.: *Multidimensional k-anonymity*, In: Technical Report 1521, University of Wisconsin. (2005)
[15] LeFevre K., DeWitt D., Ramakrishnan R.: *Workload-aware anonymization*, In: Proceeding of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2006)
[16] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," *IEEE Trans. On Knowledge and Data Engineering (TKDE)* vol. 13, no.6, pp. 1010-1027, 2001.
[17] L.Sweeney,"k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertain.Fuzz.,* vol.10,no.5,pp.557-570,2002.
[18] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," *Proc. Int'l Conf. Data Engineering (ICDE),* pp. 106115, 2007.
[19] Xiaoxun, S., Wang, H., Li, J. and Truta, T. M.: *Enhanced P-Sensitive K-Anonymity Models for Privacy Preserving Data Publishing*. In: Transaction on Data Privacy. (2008)
[20] Oliveira, S.R.M., Zaiane, O.R.: *Privacy preserving clustering by data transformation.* In: 18th Brazilian Symposium on Databases SBBD, pp. 304–318. (2003)
[21] Aggarwal, C.: *On k-Anonymity and the Curse of Dimensionality*. In: Proceeding of the 31st VLDB Conference. (2005)
[22] Bertino, E., Fovino, I.N., Provenza, L.P.: *A framework for evaluating privacy preserving data mining algorithms*. In: Data Mining and Knowledge Discovery 11(2), pp. 121–154. (2005)
[23] Li, N., Li, T. and Venkatasubramanian, S.: *t-Closeness: Privacy beyond k-anonymity and l-diversity*. In: ICDE Conference, pp. 106-115. (2007)