Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 10, October -2015

Privacy Preserving in DM using min-max normalization and noise addition

Patel Brijal H[1]. , Ankur N. Shah[2]

[1]Research Scholar, CSE Department, Parul Institute of Technology, Vadodara, India
[2]Assistant Professor, CSE Department , Parul Institute of Technology, Vadodara, India

Abstract— Privacy preserving allows sharing of privacy sensitive data for analysis purposes so it is very popular technique. so, people have ready to share their data. In recent years, privacy preserving data mining is an important one because wide availability of data is there. It is used for protecting the privacy of the critical and sensitive data and obtains more accurate results of data mining. The random noise is added to the original data in privacy preserving data Mining (PPDM) approach, which is used to publish the accurate information about original data. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data and securing the information to be misused, so that the private data and private knowledge remain as it is after mining process. This topic is used to reiterate several privacy preserving data mining technologies to protect sensitive information's privacy and obtaining data clustering with minimum information loss for multiplicative attributes in dataset.

Keywords—Privacy, K-means clustering, precision , recall, DCT.

I. INTRODUCTION

Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about them.

Generally when we talk about privacy as "keep information about me from being available to others". Our real concern is that our information not be misused. The fear is there of information to be misuse once it is released. Utilizing this distinction – ensuring that a data mining project won't enable misuse of personal information – opens opportunities that "complete privacy" would prevent. so we need technical and social solutions that ensure data will not be released. second view is corporate privacy – the release of information about a collection of data rather than an individual data item. here concerns is not about the individual's personal information; but knowing all of them enables identity theft. This collected information problem scales to large, multi-individual collections as well.

Advantages of Privacy Protection are personal information protection, proprietary or sensitive information protection, enables collaboration between different data owners (since they may be more willing or able to collaborate if they need not reveal their information), and compliance with legislative policies. Concerns about informational privacy generally relate to the manner in which personal information is collected, used and disclosed. When a business collects information without the knowledge or consent of the individual to whom the information relates, or uses that information in ways that are not known to the individual, or discloses the information without the consent of the individual, informational privacy may be violated.

Achieving the privacy of the data in data stream model is quite difficult than traditional data mining model because of the characteristics of data stream model. There is continuous flow of the data and real time processing of the data object, incorporating privacy preserving phenomenon before making data available to classification or clustering algorithm requires interrupting rate of flow incoming data, doing some processing to achieve privacy. So algorithms should be appropriately configured to cope with the interruption. Because the whole data set are not available at the same moment, achieving privacy preserving in data stream model is more difficult.

II RELATED WORK

The concept of K-anonymity is discussed in [1]. The k1-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the k-anonymity requirement, they used both generalization and suppression for data anonymization.

In general, $k$anonymity guarantees that an individual can be associated with his real tuple with a probability at most $1/k$.

In general, k anonymity guarantees that an individual can be associated with his real tuple with a probability at most $1/k$. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack. The other technique discussed in [1] is the perturbation approach, The perturbation approach works under the need that the data service is not allowed to learn or recover precise records. In the perturbation approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Another branch of privacy preserving data mining which using cryptographic techniques was developed[1]. This technique is hugely popular for two main reasons: First, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Second, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms.But, recent work has pointed that cryptography prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

The another technique in [1] is randomized response techniques. The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. The last technique in [1] is the condensation approach, which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters. This approach uses a methodology which condenses the data into multiple groups of predefined size, For each group, certain statistics are maintained. Each group has a size at least k, which is referred to as the level of that privacy-preserving approach. The greater the level, the greater the amount of privacy. At the same time, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity.

In [2] paper , the above discussed techniques are given and one more technique is discussed that is soft computing techniques. Soft computing is a consortium of methodologies which work synergistically and provides in one form or another flexible information processing capabilities for handling real-life ambiguous situations.

Soft computing techniques[2] include fuzzy logic, neural networks, genetic algorithms, and rough sets. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. It makes it possible to model imprecise and qualitative knowledge as well as the transmission and handling of uncertainty at various stages. Neural Networks are widely used for classification and rule generation. Genetic algorithms are adaptive, robust, efficient and global search methods, suitable in situations where the search space is large. Rough set is a mathematical tool for managing uncertainty that arises from indiscernibility between objects in a set. In paper [5] all this techniques are repeated with some logical modifications. The k-anonymity[5] privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least k-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the k-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a k-anonymous table through generalization and suppression remains truthful.

In the perturbation approach[5], the distribution of each data dimension is reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations.

Another branch of privacy preserving data mining which using cryptographic techniques[5] was developed. This branch became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms.

Another two approach are discussed that is randomized response techniques and condensation approach[5]. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several privacy preserving data mining technologies clearly and then proceeds to analyze the merits and shortcomings of these technologies.

III.PROPOSED WORK

Data perturbation refers to a data transformation process typically performed by the data owners before publishing their data. Two reasons of performing such data transformation. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand, the data owners want the

transformation to best preserve those meaningful data mining models which has domain-specific data properties that are critical for building, thus maintaining mining task specific data utility of the published datasets.

The data streams pre-processing stage perturb confidential data using perturbation algorithm. According to the security need, Users can flexibly adjust the data attributes to be perturbed. So, risks and threats can be effectively reduced from releasing data.

Discrete cosine transformation can preserve the Euclidian distances between original and transformed domains, which is unitary transformation. The concern of DCT is preserving privacy as well as Euclidian distances. In DCT, data characteristics remain unchanged during whole process and behavior of data as same as original data.

Proposed Algorithm:

Procedure: DCT Based Multiplicative Data Perturbation.
Input:Data Stream **D,** Sensitive attribute **S**.
Intermediate Result: Perturbed data stream **D '**.
Output: Clustering results **R** and **R'** of Data stream **D** and **D'** respectively**.**

Step 1: Given input data **D** with tuple size **n**, extract sensitive attribute $[S]_{nx3}$.
Step 2: Use Min-Max Normalization:

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Step 3: Calculate $f(x) = \sum_{n=1}^{N} \mathrm{D(n)}\left(\cos\frac{\pi(2n-1)(k-1)}{2N}\right)$

Where k = 1,2…….N

D(n) = $[S]_{nx3}$

Step 4: Multiply *f(x)* with $\frac{1}{\sqrt{n}}$ for k=1 or multiply with $\sqrt{\frac{2}{N}}$ for 2 <k <N

Step 5: Crate perturbed dataset **D'** by replacing sensitive attribute $[S]_{n\times3}$ in original dataset **D** with *f(x)***.**
Step 6: Apply ***k-Mean*** clustering algorithm with different values of k on original dataset **D** having sensitive attribute **S**.
Step 7: Apply ***k-Mean*** clustering algorithm with different values of k on perturbed dataset **D'** having perturbed sensitive attribute **P**.
Step 8: Create cluster membership matrix of results from step 5 and step 6 and analyze.

IV. IMPLEMENTATION METHODOLOGY

MOA (Massive Online Analysis) - Framework

Massive Online Analysis (MOA) [10] is a software environment for implementing algorithms and running experiments for online learning from evolving data streams. To scaling up the implementation of state of the art algorithms to real world dataset sizes, MOA GUI deals with that type of challenging problems . It contains collection of online and offline for both clustering and classification as well as tools for evaluation[10]. By getting insights into workings and problems of different approaches, Researchers getting benefits, practitioners can easily apply and compare several algorithms to real world data set and settings. MOA supports bi-directional interface with WEKA.

MOA working GUI

MOA contains data stream generators, classifiers, clustering algorithms and evaluation methods.

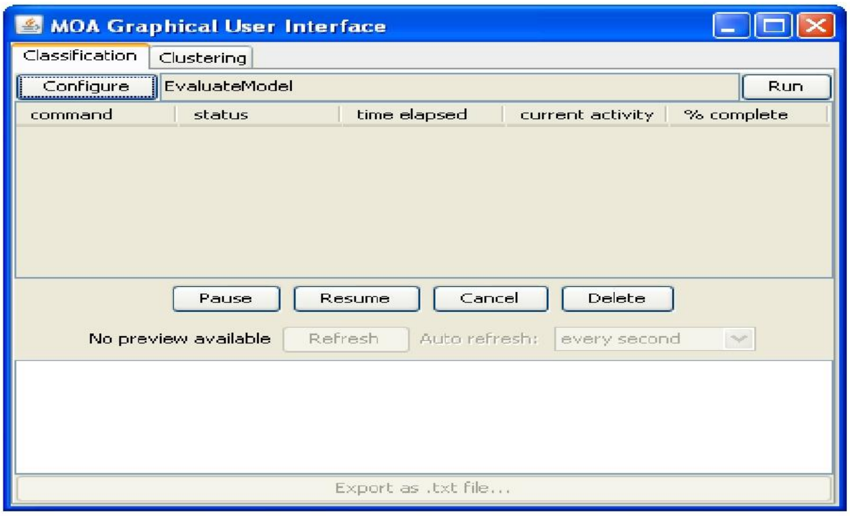

Figure 4: MOA Graphical User Interface

The Technology and platform used to implement DCT is Java and Netbeans 6.9. As input it takes no. of columns and no. of rows and extract sensitive attributes as per the domain dataset. The output is xls file of modified values of sensitive attributes.

V. EXPERIMENTAL RESULT

Results of Proposed Work
The dataset I have taken is Basketball dataset. The results are taken considering the window size and the no. of sensitive attributes. As the no. of sensitive attributes and window size decreases the accuracy increases. So the results are as follows Showing the Membership matrix for basketball dataset considering attribute age and time duration.

OriginalData set clustering		**Data clustering on dataset after data perturbation on sensitive attribute(Age,Time_Duration)**	
(All Classes)		**No. Of Correctly clustered instances**	
Clusters	**No. instances in each Cluster**	**Age**	**Time_Duration**
C1	5681	5239	5315
C2	14131	11348	13275
C3	10188	8195	8878
Total	30000	24782	27468
Result	**Accuracy (%)**	**82.60**	**91.56**

Figure 5: Membership Matrix for Basketball Dataset (w=3000 and k=3)

Original Data set clustering		**Data clustering on dataset after data perturbation on sensitive attribute(Age,Time_Duration)**	
(All Classes)		**No. Of Correctly clustered instances**	
Clusters	**No. instances cover in each Cluster**	Age	Time_Duration
C1	4112	3129	4007
C2	7790	6793	7684
C3	8407	8305	8308
C4	2730	2001	2649
C5	4961	3212	3569
Total	28000	23441	26216
Result	**Accuracy (%)**	**83.71**	**93.62**

Figure 6 :Membership Matrix for Basketball Dataset (w=2000 and k=5)

OriginalData set clustering		**Data clustering on dataset after data perturbation on sensitive attribute(Time_Duration, Duration)**	
(All Classes)		**No. Of Correctly clustered instances**	
Clusters	**No. instances cover in each Cluster**	Age	Time_Duration

C1	1216	1147	1201
C2	25746	23774	25607
C3	1038	956	1029
Total	28000	25877	27837
Result	**Accuracy (%)**	92.41	99.41

Figure 7: Membership Matrix for Basketball Dataset (w=2000 and k=3)

Now the accuracy results of attribute age and time duration are as showing in the graph based on Precision and recall measure which are used to measure the accuracy.

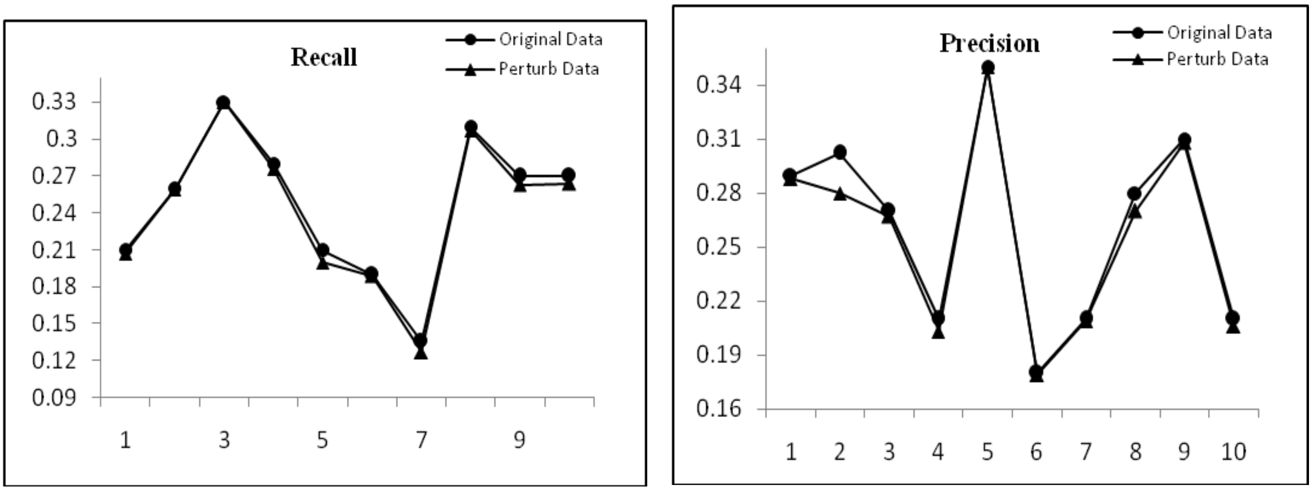

Figure 8: Accuracy on attribute Time_Duration in Basketball dataset(w=3000 and k=3)

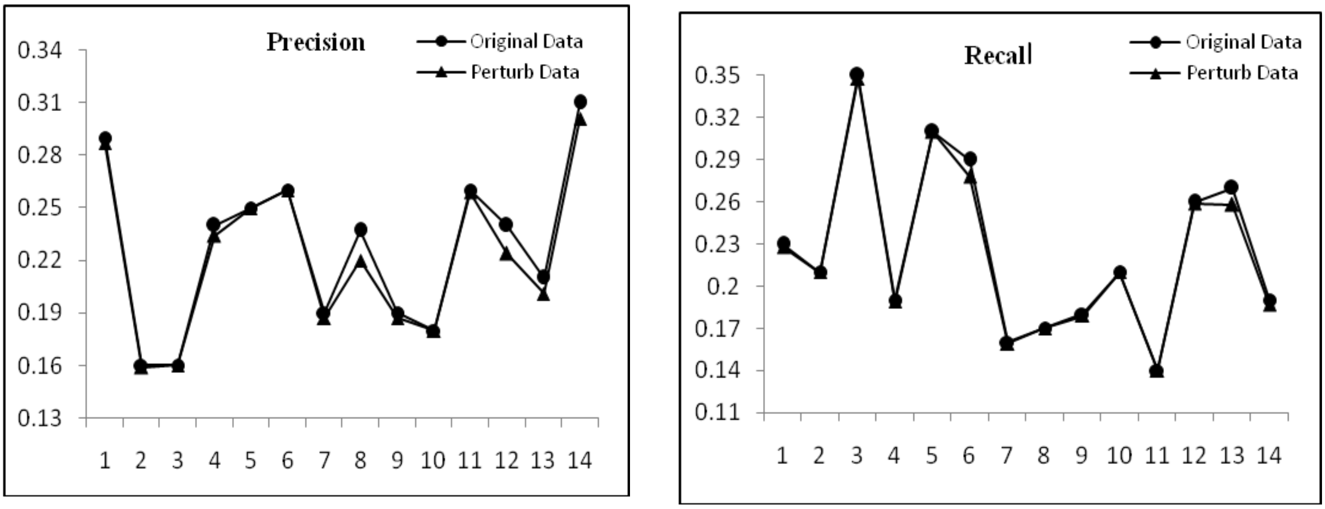

Figure : 9 Accuracy on attribute Time_Duration in Basketball dataset (w=2000 and k=3)

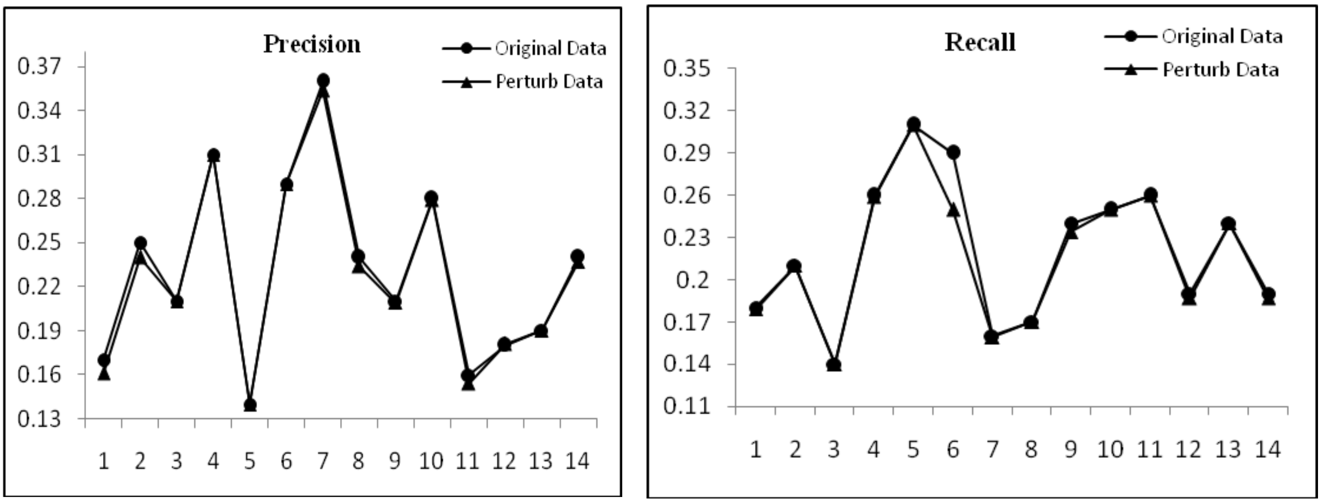

Figure 10 : Accuracy on attribute Time_Duration in Basketball dataset (w=2000 and k=5)

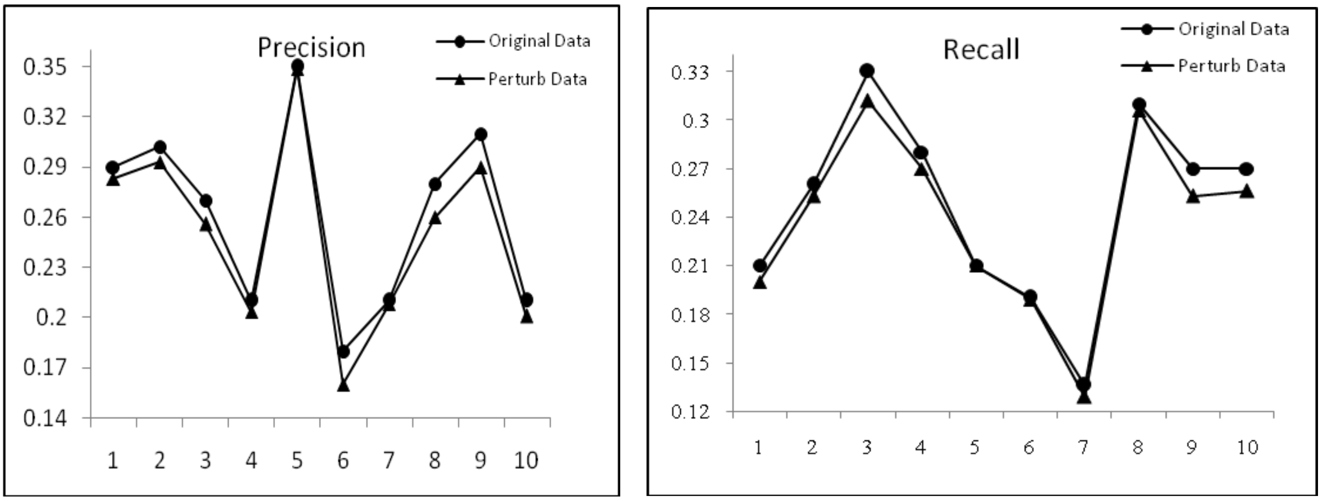

Figure 11 : Accuracy on attribute Age in Basketball dataset (w=3000 and k=3)

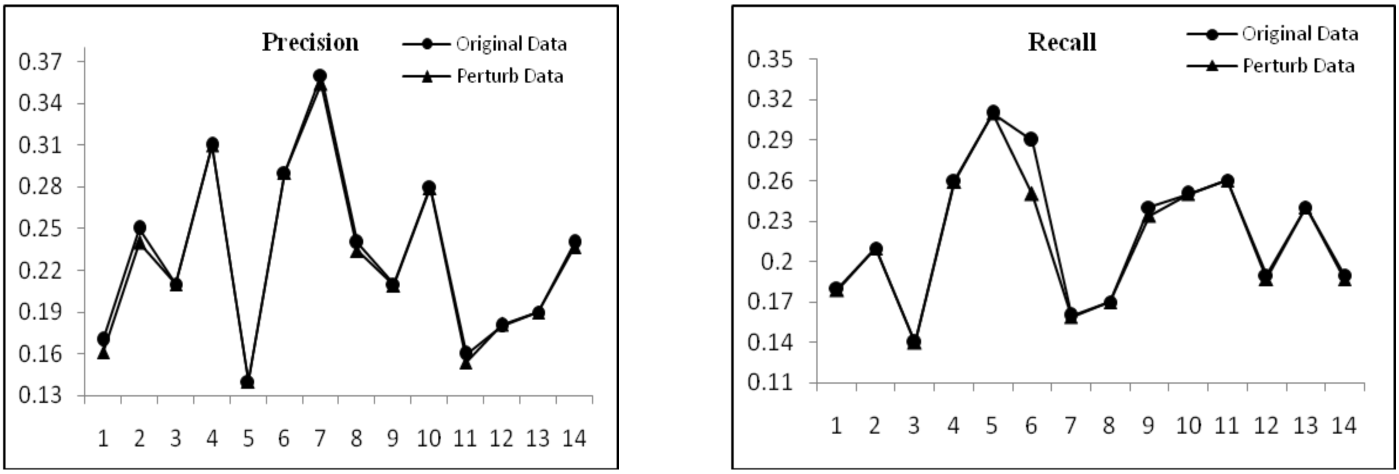

Figure 12 : Accuracy on attribute Age in Basketball dataset(w=2000 and k=5)

As shown in results , if the window size is 2000 and no. of sensitive attribute is 3, than there is higher accuracy of the data that is 99.41%. Like this all the sensitive attributes results should taken.

VI. CONCLUSION

Privacy is the most important approach to protect the sensitive data. The data which they don't want to share, People are very much worried about their sensitive information. My survey in this topic focuses on the existing techniques which have been already present in the field of Privacy Preserving Data Mining. From my analysis, I have found that, no single technique that is used in all domains. All techniques perform in a different way as the type of data as well as the type of application or domain. But still from my analysis, I can conclude that Random Data Perturbation and Cryptography techniques perform better than the other existing methods. Cryptography is best technique for encrypting sensitive information. On the other hand Data Perturbation will help to preserve data so sensitive information is maintained. And at last, I want to say that perturbation technique with normalization is used to improve the level of privacy so perturbation technique with normalization is more accurate than all other existing techniques.

ACKNOWLEDGMENT

I thanks to Assistant Professor Mr. Ankur N. Shah who is from Computer & Science Department of Parul Institute And Technology .he is also my guide who inspire me for doing such work and due to their support here I am able to write this review paper and she also provide me good knowledge about perturbation techniques and Guide me for doing this work.

REFERENCES

RESEARCH PAPERS

[1] W.T. Chembian1, Dr. J.Janet, "*A Survey on Privacy Preserving Data Mining Approaches and Techniques*", Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010 Chennai, India.

[2] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "*Privacy Preserving Data Mining Techniques: Current Scenario and future prospects* ", 2012 Third International Conference on Computer and Communication Technology

[3]Kiran Patel, Hitesh Patel, Parin Patel, "*Privacy Preserving in Data stream classification using different proposed Perturbation Methods*", © 2014 IJEDR | Volume 2, Issue 2 | ISSN: 2321-9939

[4] T.J. Trambadiya,and P. bhanodia , "*A Heuristic Approach to Preserve Privacy in Stream Data with Classification*", International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 1, pp.1096-1103, Jan -Feb 2013.

[5] Jian Wang ,Yongcheng Luo, Yan Zhao Jiajin Le," *A Survey on Privacy Preserving Data Mining*", First International Workshop on Database Technology and Applications, 978-0-7695-3604-0/09 © 2009 IEEE

[6]R.Vidya Banu, N.Nagaveni," *Preservation of Data Privacy using PCA based Transformation*", International Conference on Advances in Recent Technologies in Communication and Computing, © 2009 IEEE

[7]Hitesh Chhinkaniwala1 and Sanjay Garg," *Tuple Value Based Multiplicative Data Perturbation Approach To Preserve Privacy In Data Stream Mining*", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013

[8] K. Wankhade ,T.Hasan and R.thool "A Survey: Approaches for Handling Evolving Data Streams", International Conference on Communication Systems and Network Technologies IEEE, pp. 621-625, 2013

[9]S.Guha, A.Meyerson, N.Mishra and R.Motvani "Clustering Data Streams: Theory and Practice", IEEE transactions on knowledge and data engineering, Vol. 15, NO. 3, pp. 515- 528, MAY/JUNE 2003

[10] M. Naga lakshmi1, K Sandhya Rani, "Privacy Preserving Clustering Based on Discrete Cosine Transformation" ,IJIRSET,Vol. 2, Issue 9, September 2013

[11] A.Bifet, R.Kirkbry, P. Kranen and P. Reutemann ," MOA: Massive Online Analysis manual", March 2012

BOOKS

[B1] "DATA MINING" concepts and techniques by Jiawei Han, Michleline Kamber and Jian pei Third Edition . ELSEVIER , Morgan Kaufmann publisher

Author Details

Patel Brijal H. received the B.E. degree in Information Technology from Sigma Institute of Engineering, Gujarat Technological University in 2013.

Currently she is doing her M.E. Degree from Parul Institute of Technology of Gujarat Technological University and her interesting area of research is in Privacy Preserving of Data Mining.