

Design and Development of Privacy Preserving
Techniques for Data Stream Mining

A Thesis Submitted To

Nirma University

In Partial Fulfillment Of The Requirements For

The Degree Of

Doctor Of Philosophy

In

Technology & Engineering

By

Mr Solanki Pareshkumar Mahendrabhai
(11EXTPHDE58)



Institute of Technology

Nirma University

Ahmedabad-382481

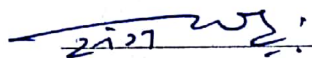
Gujarat, India

December, 2018

Nirma University Institute of Technology Certificate

This is to certify that the thesis entitled Design and Development of privacy preserving Techniques for Data Stream Mining has been prepared by Mr Solanki Pareshkumar Mahendrabhai under my supervision and guidance. The thesis is his original work completed after careful research and investigation. The work of the thesis is of the standard expected of a candidate for Ph.D. Programme in Computer Science and Engineering and I recommend that it be sent for evaluation.

Date:
16.07.2018



Signature of the Guide


Forwarded Through:


Dr. Madhuri Bhansale

(i) Name and Signature of the Head of the Department



(ii) Name and Signature of the Dean Faculty of Technology & Engineering


17.7.18

(iii) Name and Signature of the Dean Faculty of Doctoral Studies and Research

To

Executive Registrar

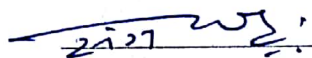
Nirma University


17.7.18

Nirma University Institute of Technology Certificate

This is to certify that the thesis entitled Design and Development of privacy preserving Techniques for Data Stream Mining has been prepared by Mr Solanki Pareshkumar Mahendrabhai under my supervision and guidance. The thesis is his original work completed after careful research and investigation. The work of the thesis is of the standard expected of a candidate for Ph.D. Programme in Computer Science and Engineering and I recommend that it be sent for evaluation.

Date:
16.07.2018




Signature of the Guide


Forwarded Through:

 Dr. Madhuri Bhansale

(i) Name and Signature of the Head of the Department

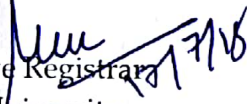


(ii) Name and Signature of the Dean Faculty of Technology & Engineering


17.7.18

(iii) Name and Signature of the Dean Faculty of Doctoral Studies and Research

To
Executive Registrar
Nirma University



Nirma University Institute of Technology Certificate

This is to certify that the thesis entitled Design and Development of privacy preserving Techniques for Data Stream Mining has been prepared by Mr Solanki Pareshkumar Mahendrabhai under my supervision and guidance. The thesis is his original work completed after careful research and investigation. The work of the thesis is of the standard expected of a candidate for Ph.D. Programme in Computer Science and Engineering and I recommend that it be sent for evaluation.

Date:

Signature of the Guide

Forwarded Through:

(i) Name and Signature of the Head of the Department

(ii) Name and Signature of the Dean Faculty of Technology & Engineering

(iii) Name and Signature of the Dean Faculty of Doctoral Studies and Research

To
Executive Registrar
Nirma University

Nirma University Institute of Technology Declaration

I, Mr Solanki Pareshkumar Mahendrabhai, registered as Research Scholar, bearing Registration No.11EXTPHDE58 for Doctoral Programme under the Faculty of Technology & Engineering of Nirma University do hereby declare that I have completed the course work, pre-synopsis seminar and my research work as prescribed under R.Ph.D. 3.5.

I do hereby declare that the thesis submitted is original and is the outcome of the independent investigations / research carried out by me and contains no plagiarism. The research is leading to the discovery of new techniques already known. This work has not been submitted by any other University or Body in quest of a degree, diploma or any other kind of academic award.

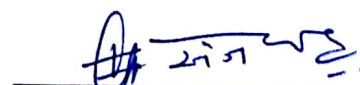
I do hereby further declare that the text, diagrams or any other material taken from other sources (including but not limited to books, journals and web) have been acknowledged, referred and cited to the best of my knowledge and understanding.

Date: 16/07/2018



Signature of Student

we endorse the above declaration made by the student.


Dr Sanjay Gang

Name & Signature of the Guide

Nirma University Institute of Technology Declaration

I, Mr Solanki Pareshkumar Mahendrabhai, registered as Research Scholar, bearing Registration No.11EXTPHDE58 for Doctoral Programme under the Faculty of Technology & Engineering of Nirma University do hereby declare that I have completed the course work, pre-synopsis seminar and my research work as prescribed under R.Ph.D. 3.5.

I do hereby declare that the thesis submitted is original and is the outcome of the independent investigations / research carried out by me and contains no plagiarism. The research is leading to the discovery of new techniques already known. This work has not been submitted by any other University or Body in quest of a degree, diploma or any other kind of academic award.

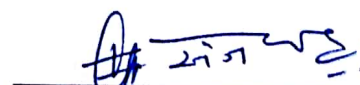
I do hereby further declare that the text, diagrams or any other material taken from other sources (including but not limited to books, journals and web) have been acknowledged, referred and cited to the best of my knowledge and understanding.

Date: 16/07/2018



Signature of Student

we endorse the above declaration made by the student.


Dr Sanjay Gang

Name & Signature of the Guide

Nirma University

Institute of Technology

Declaration

I, Mr Solanki Pareshkumar Mahendrabhai, registered as Research Scholar, bearing Registration No.11EXTPHDE58 for Doctoral Programme under the Faculty of Technology & Engineering of Nirma University do hereby declare that I have completed the course work, pre-synopsis seminar and my research work as prescribed under R.Ph.D. 3.5.

I do hereby declare that the thesis submitted is original and is the outcome of the independent investigations / research carried out by me and contains no plagiarism. The research is leading to the discovery of new techniques already known. This work has not been submitted by any other University or Body in quest of a degree, diploma or any other kind of academic award.

I do hereby further declare that the text, diagrams or any other material taken from other sources (including but not limited to books, journals and web) have been acknowledged, referred and cited to the best of my knowledge and understanding.

Date:

Signature of Student

we endorse the above declaration made by the student.

Name & Signature of the Guide

Publications related to the thesis

- Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Privacy Preserving in Data Stream Mining Using Multi-iterative K-Anonymization*, International Journal of Data Mining and Emerging Technologies, Volume 8, PP. 1-9, 2017.
- Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Heuristic-based hybrid privacy preserving data stream mining approach using SD-Perturbation and Multi-Iterative K-Anonymization*, International Journal of Knowledge Engineering and Data Mining, Inderscience, Volume 5, No. 4, PP. 306-332, 2018.
- Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *A Comprehensive Review of Privacy Preserving Techniques in Data Mining and Data Stream Mining*, International Journal of Data Mining and Emerging Technologies, Volume 8, PP. 105-116, 2018.
- Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Privacy Preserving Data Stream Mining using Hybrid Geometric Data Perturbation*, International Journal of Research in Electronics and Computer Engineering, Volume 5, PP. 107-116, 2017.
- Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *SD-Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering*, International Journal of Information and Communication Technology, inderscience, 2018. (Paper in Communication)

Abstract

Data mining is the crucial field of pulling out information from bulky dataset with diverse application areas such as healthcare, banking and financial, telecommunication, shopping records, personal data and so on. These applications frequently produce huge volume of data which is stored statically and dynamically in the available network. The mined statistics can be in the form of clusters, patterns, rules and classification. Distribution of such data is demonstrated to be advantageous for data mining application. This dataset frequently encompasses classifiable information individually and consequently freeing such data may result in privacy breaches. Preserving privacy while delivering data is a fundamental study area in data security and also it is a major issue in delivering individual exact sensitive information. Efficient preservation of data proprietor's privacy is a crucial issue while broadcasting the data for analysis purpose. As per our knowledge, dataset is an essential asset for industry in order to take a decision by examining it. In order to distribute the data along side preserving privacy, the data proprietor must come up with a result which accomplishes the double goal of privacy preservation as well as accuracy of data mining task, mostly clustering and classification. Data mining can be valuable in many applications, but due to insufficient protection the data may be abused for other goals. It is essential to prevent revealing of not only the individual confidential information but also the critical knowledge. Generally, data proprietors do not find it safe to publish datasets for mining purpose because of their worry that releasing of data may compromise an individual's private information. Perturbation and Anonymizing datasets before releasing overcomes such a fear as it guarantees secrecy of personal information. But, protecting personal information and achieving mining results as close as that of with original datasets poses great challenges. The Proposed research

work tries to find out solutions for this growing concern. Several algorithms have been proposed that understand the characteristics of the dataset and perturb either sensitive attribute values or keep sensitive attribute's values unchanged and anonymized quasi-identifier's values. Various data perturbation and anonymization based algorithms proposed so far have focused mainly on static data and very few are on data streams. Heuristic based data perturbation has been proposed where privacy has been maximized through computed tuple values for each instance and user define sensitive drift with minimum information loss. Proposed algorithm has been evaluated to measure information gain and to achieve privacy. Many datasets contain multiple sensitive attributes so, there is a need to provide perturbation and anonymization to preserve the privacy. Based on this concern, the research work is also carried out for detail analysis of data anonymization alternatives and proposed heuristic based PRIVACYearn based multi-iterative k-anonymization and perturbation approach in data stream. This approach also proposes to find out the best fit generalization that leads to minimum loss of information and better protection of individual's privacy. Finally, we have proposed heuristic based geometric data perturbation in data stream. Developed algorithms for data perturbation and anonymization have been tested using wide range of standard datasets over frequently used mining algorithms like, K-Mean clustering and Naive Bayes classification.

Acknowledgements

Nirma University gave me the best opportunity to grow and develop the researcher in me and find my strength to face and survive the competition with excellence. First of all, I would like to thank and be grateful to Nirma University and all kind of research facility which was provided to me. I would like to thank Dr Sanjay Garg, Guide and the Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for motivating me and providing continuous support throughout my Doctoral studies. I would also like to thank Dr. Hitesh Chhikaniwala for providing continuous guidance throughout my Doctoral studies.

My sincere thanks to the reviewers Dr D. C. Jinwala and Dr Minal Bhise, who had given their valuable feedback during my Research Progress Committee meetings.

I am also thankful to all faculty members, staff members of the department for providing all types of resources and support in time.

There are no words to thank the Almighty for gifting me with wonderful children DIVYAM, who bore the wrath of my journey towards this research. My wife DHARAL needs special accolade for her patience and constant support provided throughout this work. Finally, my Sincerest thanks to the most important and special people of my life - My Parents; who have been ever motivating, endearing and highly cooperative in all my endeavors.

I thank one and all, who have kept encouraging and motivating me. Without everyone's support, this thesis is beyond imagination.

Mr Paresh Solanki

11EXTPHDE58

Contents

Certificate	i
Declaration	ii
Abstract	v
Acknowledgements	vii
List of Tables	xiii
List of Figures	xv
Abbreviations	xix
1 Introduction	1
1.1 Privacy Preserving in Data Mining (PPDM)	3
1.1.1 Privacy issues in data mining	4
1.1.2 Privacy Preserving in Data Stream Mining (PPDSM)	6
1.2 Motivation	7
1.3 Objectives	8
1.4 Scope of the Research Work	8
1.5 Major contributions	9
1.6 Organization of Thesis	10
2 Background	11
2.1 Data Mining	11
2.2 Data Stream Model	12
2.3 Data Stream Mining	12
2.3.1 Data Stream Classification	14
2.3.2 Data Stream Clustering	16
2.4 Data Perturbation	19
2.5 Quasi-Identifiers (<i>QI</i>)	19
2.6 K-Anonymity	20
2.7 MOA-Massive Online Analysis	21
2.7.1 Characteristics	21
2.7.2 MOA-Features	22
2.7.3 Evaluation Measures	23

3	Related work	27
3.1	Heuristic Based Methods	31
3.1.1	k-Anonymity	32
3.1.2	Personalized privacy preservation	33
3.1.3	Privacy preservation based on utility	33
3.2	Cryptographic Based Methods	34
3.3	Privacy Preserving Data Stream Mining (PPDSM)	35
3.4	Summary of Major Research Contributions	37
3.5	Research Issues & Challenges	44
4	Heuristic based Privacy Preserving Data Stream Mining using Hybrid Geometric Data Perturbation	47
4.1	Problem formulation	47
4.2	Proposed framework	48
4.3	Proposed Algorithm	49
4.4	Performance Evaluation	52
4.4.1	Experimental setup	53
4.4.2	Experimental Results	55
4.4.3	Comparison of Proposed Approach with GDP Approach . . .	71
4.5	Summary	71
5	SD – Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering	73
5.1	Problem formulation	74
5.1.1	Privacy in Data Stream Mining	74
5.1.2	Privacy in Data Mining	74
5.2	Proposed framework	75
5.2.1	Formal Analysis of proposed work	78
5.3	Proposed Algorithm	80
5.4	Performance Evaluation	81
5.4.1	Experimental setup	81
5.4.2	Experimental Results	83
5.4.3	Cluster Membership Matrix (CMM)	83
5.4.4	Precision & Recall Measurement	87
5.4.5	Privacy Gain Measurement	91
5.5	Summary	93
6	Heuristic based hybrid privacy preserving data stream mining approach using Perturbation and K-anonymization	95
6.1	Problem Formulation	98
6.2	Proposed framework	99
6.3	Proposed Algorithm	101
6.4	Performance Evaluation	110
6.4.1	Comparison of proposed approach with K-Anonymization . .	119
6.5	Summary	121

7	Conclusion and Future Scope	123
7.1	Conclusion	123
7.2	Limitations	125
7.3	Future Scope	125
A	Test cases with standard data sets applied over proposed algorithms	127
	Index	131
	Bibliography	133

List of Tables

2.1	Medical dataset	21
2.2	Anonymized dataset (2-Anonymity)	21
3.1	Heuristic-based techniques	37
3.2	Reconstruction-based techniques	39
3.3	Cryptography-based techniques	42
3.4	Privacy preserving data stream mining	43
4.1	Clustering Membership Matrix(CMM)	55
4.2	Information Gain for Covertypes Dataset (W=1000, Angle = 30°, Sequence = TDP_SDP_RDP)	56
4.3	Information Gain for Covertypes Dataset (w=3000, Angle = 30°, Sequence = TDP_SDP_RDP)	57
4.4	Information Gain for Covertypes Dataset (w=5000, Angle = 60°, Sequence = RDP_SDP_TDP)	57
4.5	Information Gain (Covertypes Dataset)	58
5.1	Description about datasets	83
5.2	Clustering Membership Matrix (CMM)	84
5.3	K-Means cluster result of the perturbed datasets for window size=50 & Sensitive Drift=10%	88
5.4	Parametric analysis on actual data values vs perturbed data values	89
5.5	Bias in Mean (BIM) and Bias in Standard Deviation (BISD)	91
6.1	Quasi-identifiers in Bank Marketing data set with applied generalization level for k-anonymization	112
6.2	Quasi-identifiers in Adult data set with applied generalization level for k-anonymization	112
6.3	Privacy Gain result of Bank Marketing Data set (k=3, w=3000)	115
6.4	Privacy Gain result of Adult and Bank Marketing Dataset (W=10000)	115
6.5	Performance of PRIVACYearn based Multi-iterative k-Anonymization algorithm (window size=3000)	118
6.6	Performance of PRIVACYearn based Multi-iterative k-Anonymization algorithm (window size=10000)	118
6.7	Comparison of proposed approach with existing k-anonymization approach (Privacy outcomes)	119
6.8	Comparison of proposed approach with existing k-anonymization approach (Execution Time)	120

List of Figures

2.1	The general process of data stream mining	14
2.2	Round of Data Stream Classification	14
2.3	MOA data stream clustering setup	24
2.4	MOA visualization components	25
3.1	Privacy Preserving Data Mining Techniques	30
3.2	Data Hiding & Rule Hiding based PPDM Techniques	31
4.1	Framework of Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation	49
4.2	Flow of the proposed Framework	52
4.3	Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 1000)	59
4.4	Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 1000)	59
4.5	Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, w = 1000)	60
4.6	Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, w = 1000)	60
4.7	Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, w = 1000)	60
4.8	Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 3000)	61
4.9	Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 3000)	61
4.10	Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, w = 3000)	62
4.11	Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, w = 3000)	62
4.12	Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, w = 3000)	63
4.13	Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 5000)	63
4.14	Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 5000)	64
4.15	Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, w = 5000)	64

4.16	Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, w = 5000)	65
4.17	Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, w = 5000)	65
4.18	Accuracy on Covertypes Dataset (w = 1000, Angle = 30, Sequence = SDP_TDP_RDP)	66
4.19	Accuracy on Covertypes Dataset (w = 1000, Angle = 30, Sequence = SDP_TDP_RDP)	66
4.20	Accuracy on Covertypes Dataset (w = 5000, Angle = 60, Sequence = SDP_TDP_RDP)	67
4.21	Accuracy on Covertypes Dataset (w = 5000, Angle = 60, Sequence = SDP_TDP_RDP)	67
4.22	Accuracy on Bank Dataset (w = 1000, Angle = 60, Sequence = SDP_TDP_RDP)	68
4.23	Accuracy on Bank Dataset (w = 1000, Angle = 60, Sequence = SDP_TDP_RDP)	68
4.24	Accuracy on Electricity Dataset (w = 1000, Angle = 30, Sequence = SDP_TDP_RDP)	69
4.25	Accuracy on Electricity Dataset (w = 1000, Angle = 30, Sequence = SDP_TDP_RDP)	69
4.26	Accuracy on Electricity Dataset (w = 3000, Angle = 60, Sequence = SDP_TDP_RDP)	70
4.27	Accuracy on Electricity Dataset (w = 3000, Angle = 60, Sequence = SDP_TDP_RDP)	70
4.28	Comparison of proposed approach with GDP	71
5.1	Identical independently distributed noise addition into original data streams	75
5.2	Principal Component Analysis (PCA) based data reconstruction . . .	75
5.3	Entire process of SD Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering . .	76
5.4	Extended framework of Massive Online Analysis (MOA) using SD-Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering	78
5.5	Accuracy obtained for Covertypes Dataset (Elevation)	85
5.6	Accuracy obtained for Covertypes Dataset (Aspect)	85
5.7	Accuracy obtained for Covertypes Dataset (Slope)	86
5.8	Accuracy obtained for ElectricNorm Dataset (swPrice)	86
5.9	Accuracy obtained for ElectricNorm Dataset (swDemand)	86
5.10	Accuracy obtained for Bank Marketing Dataset (Balance)	87
5.11	Accuracy obtained for Bank Marketing Dataset (Duration)	87
5.12	Accuracy measurement of Cover type dataset (Sensitive Attributes = (Aspect, Slope,Elevation) window size = 50, Sensitive drift = 10%) .	90
5.13	Accuracy measurement of Data Set Electric Norm (Sensitive Attributes = Demand, Price, window size=50, Sensitive drift=10%)	90
5.14	Accuracy measurement of Data Set Bank Marketing (Sensitive Attributes = Balance, Duration, window size=50, Sensitive drift=10%) .	91
5.15	Privacy Measurement using Variance Difference	92

6.1	Flow chart of the proposed work	98
6.2	Framework of Heuristic based hybrid privacy preserving data mining approach using Perturbation and K-anonymization	100
6.3	Medical Dataset	104
6.4	Anonymized Dataset ($Gender_1$), Privacy Gain=0%	105
6.5	Anonymized Dataset ($Zipcode_1$), Privacy Gain=0%	106
6.6	Anonymized Dataset (Age_1), Privacy Gain=15%	106
6.7	Anonymized Dataset ($Age_1, Gender_1$), Privacy Gain=60%	107
6.8	Anonymized Dataset ($Age_2, Gender_1$), Privacy Gain=60%	107
6.9	Anonymized Dataset ($Age_1, Gender_1, Zipcode_1$), Privacy Gain=60%	108
6.10	Anonymized Dataset ($Age_2, Gender_1, Zipcode_1$), Privacy Gain=60%	108
6.11	Anonymized Dataset ($Age_1, Gender_1, Zipcode_2$), Privacy Gain=60%	109
6.12	Anonymized Dataset ($Age_1, Gender_1, Zipcode_3$), Privacy Gain=60%	109
6.13	Anonymized Dataset ($Age_2, Gender_1, Zipcode_4$), Privacy Gain=90%	110
6.14	Accuracy measured using Naive Bayes algorithm on Adult and Bank Marketing Dataset	114
6.15	Privacy Gain of bank marketing data set	116
6.16	Privacy Gain of Adult data set	116
6.17	Privacy gain outcome of bank dataset (window size=10000)	117
6.18	Privacy gain outcome of bank dataset (window size=5000)	117
6.19	Comparison of proposed approach with k-Anonymization	120

List of Abbreviations

DM	Data Mining
DSM	Data Stream Mining
PPDM	Privacy Preserving Data Mining
PPDSM	Privacy Preserving Data Stream Mining
BIS	Bias in Mean
BISD	Bias in Standard Deviation
OLAP	On Line Analytical Process
KDD	Knowledge Discovery in Databases
ML	Machine Learning
NN	Neural Networks
DSMS	Data Stream Management System
VFDT	Very Fast Decision Tree Learner
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchie
CU	Category Utility function
QI	Quasi Identifiers
MOA	Massive Online Analysis
GUI	Graphical User Interface
SSN	Social Security Number
SMC	Secure Multiparty Computation
DGH	Domain Generalization Hierarchy
TDS	Top Down Specialization
KIDS	K-anonymIzation DataStreambased on sliding window
CASTLE	...	Continuously Anonymizing STraming data via adaptive cLustEring
PCDS	Privacy Preserving Classification of Data Stream
DSPS	Data Streams Preprocessing System
ODMS	Online Data Mining System
WASW	Weighted Average Slide Window
ASD	Average Squared Distance
EM	Expectation Maximization
SWAF	Sliding Window Anonymization of data stream Framework
GDP	Geometric Data Perturbation
CMM	Cluster Membership Matrix
TDP	Translation Data Perturbation
SDP	Scaling Data Perturbation
RDP	Rotation Data Perturbation
PCA	Principal Component Analysis
SD	Sensitive Drift

Chapter 1

Introduction

In today's era of advanced technology, online services are widely used in social networks, shopping sites or supermarket transaction data, medical services, bank card usage records, mobile or telephone call records, statistical data of government sectors, astronomical records, Patient's records in hospitals, and other online web portal data. The evolution of huge databases is experiencing upward trend because of advancement in digital data procurement and repository technology. The data collected thus are further used for the analysis and research purposes. The discipline evolved by means of this advancements is known as DM(Data Mining). (Data Mining). Data Mining is the way of pulling out statistics from the enormous datasets using the procedures and methods design from the arena of ML (Machine Learning), statistics and DBMS (Data Base Management System) (Feelders, Daniels, and Holsheimer).

The procedure of data mining is not only to examine datasets but also discover unexpected relationships and sum up the data in such an innovative way that it is understandable as well as beneficial to the data proprietor. Summaries and relationships which are retrieved over a data mining are frequently indicated as models or forms. Instances contain rules, graphs, clusters, tree structures, and recurrent forms in time series and others (David J Hand). Data mining, generally called as information finding in huge data, permits companies and administrations in making premeditated conclusions by gathering, examining and retrieving commercial data. It uses the diversity of applications like Reporting and query applications, Decision Support System applications and Analytical applications.

Nowadays, Streaming Data can be understood as a nonstop and varying sequence of data that constantly comes to a work-station to process or store (Malik, Ghazi, and Ali). As an example, satellite remote sensing systems continuously generates data. Generated data is enormous, fast-changing, sequentially ordered, and possibly unbounded. Such, characteristics create stiff challenge in the area of data streams. Data stream mining which is known as knowledge-structure pulls out as patterns and models from endless streaming data. Data streams have diverse tasks in numerous aspects, like mining, storage, querying and computational. For the reason of streaming data necessities, it is vital to create the latest procedures to replace the old ones.

Main challenges in data stream mining are: for streaming data, creating fast data mining methods and essentially identify promptly changing concepts and scattering because of enormously varying nature of streaming data (Aggarwal) (Chu). Some challenges of data stream mining are unpractical to store the entire data, Random access is costly, easy calculation per data because of time and space limitations. The quick development in online technology and communications technology has led to the emergence of streaming data. For the reason, that of sole properties of data streams, the study of data mining methods has moved from existing static data mining to data stream mining. Data stream mining is takeout information form non-stop flow of data ((Chang and Lee))((Golab and Özsu)). In data stream processing, we assume that training instances can be concisely studied a single time only because data stream reach at high speed and then after must be thrown out to create a room for next streaming data. The algorithm has no control to quickly process such streaming data instances and it is necessary to update its structure incrementally as every instance is observed. We require a supplementary desirable property where the model is prepared to be applied at any point between training samples (Chaudhry, Shaw, and Abdelguerfi).

Existing traditional data mining procedures which are used in applications that produce learning models are static in nature because persistent data is available. Since, entire data is present before we make it available to our learning procedure, statistical information of the data distribution can be known in well in advance. If the existing bigger dataset are sampled in order to accommodate with

the memory which is smaller in size with respect to bigger size of data, then incremental process on data is not possible. The learning model starts each single sample processing from the scratch. There is no mode of intermediate investigation of the outcome. Currently, in the area of data processing, tools are not appropriate for given data model (Golab and Özsu) because data arises in the form of flows (streams). Reason of sole belongings of data stream, It is difficult to regulate the arriving order at place and also impossible to store whole stream in memory. Similarly, execution over streams runs endlessly and incrementally returns fresh outcomes as latest data arrives.

Streaming data or static datasets are helpful for modernization and enhancements of the user experience, but its treatment also offers challenges to individual and business privacy. This thesis investigates the trade-offs among utility and privacy that arises when online amenities accumulate, data-mine and distribute user or business data, and make algorithms that can enable the services to stabilize those trade-offs. The main objective of our thesis is to develop techniques that data owners can use to process sensitive data for protecting sensitive information and guarantee information functionality within an acceptable boundary.

1.1 Privacy Preserving in Data Mining (PPDM)

Privacy Preserving Data Mining (PPDM) is recognized as the novel period of study in data mining. PPDM algorithms deal with extraction of hidden predictive information from large databases and preserve sensitive information from divulgence or inference. Three philosophical approaches used in PPDM research are: (1) Data hiding, (2) Rule hiding, (3) Secure Multiparty Computation (SMC). In order to preserve privacy, various data transformation methods are used for privacy computations. Various methods for privacy preserving have been proposed; still more research work is being carried out. The main objective is to preserve confidentiality of data, as extracting important information from large databases is achieved by data mining.

The key purpose of this technique is safety of dataset and keeping the usefulness and certainty of mined rules at uppermost level. There are two folds of the central attention of the PPDM : 1. Sensitive attribute data such as name, ad-

dress, ID and such others should be trimmed or changed out from the real records so it is not possible to opponent to reveal user's privacy. 2. Sensitive knowledge which can be mined from data set with the help of data mining methods should similarly be omitted, Since such information can likewise sound compromise data privacy. The aim of PPDM is to establish the procedures for changing the real data in certain manner, so that, the sensitive data and the sensitive information remain private after data mining.

1.1.1 Privacy issues in data mining

currently, we are having number of e-devices through which we perform several doings such as send emails, do credit cards transactions, security cards swapping, talking over phones or mobiles etc. Preferably, the data should be produced and composed with the consent of the data focuses. The accumulators should provide some guarantee that the individual privacy will be protected. Alternatively, the subordinate use of composed data is furthermore very general. It is a general exercise that governments or private sectors sell the composed data to other private sectors and government sectors. These sectors use this composed data for their personal purposes. Currently, mining of data is an extensively recognized method for vast variety of sectors. The returns should be acknowledged and should not be overvalued.

The entire method of data mining ranges from collecting records to finding the information, which normally holds sensitive individual information. Frequently catch uncovered to numerous gatherings including accumulators, proprietors, operators and miners. Expose of such type of sensitive information can lead to crack of individual privacy. For instance, credit-card data set may reveal an individuals hidden style (way of life) with adequate accuracy. Hidden data can similarly be revealed by connecting various data sets belonging to massive data repositories (Fienberg) and retrieving web log data (Thuraisingham). An intruder or opponent can mug up sensitive attribute values such as types of disease (e.g. Heart Problem), and salary information (e.g. AUD 82,000) of a particular individual, over and done with re-identification of the data from an uncovered dataset. Here, it is pointed out that removing the person-name and additional identifiers

like, SSN (Social Security Number), Driving license Number may not guarantee that the privacy preservation is achieved of individual, because privacy preserving in Data Stream Mining or data mining frequently be individually recognized from the mixture of other attributes. So, if he/she has adequate further knowledge about an individual, then it is very easy for opponent to be able to re-recognize a record from a dataset. If the intruder has enough knowledge such as past-life information, religion information, married status and total of kids of the individual then there is an emergent concern about delicate individual information being free which would be exploited.

Additional private information, even though not sensitive as like medical records, can similarly be measured to be private and susceptible to malicious manipulation. Instance like, credit-card data, book details and DVD details and phone or mobile call details made by an individual can be used to observe his/her individual lifestyles. Community concern is mostly triggered by the so-called subordinate use of individual information without permission of the subject. In other words, users feel powerfully that their individual information should not be wholesaled to other sectors without their prior consent. We know that there are massive benefits of Data Mining (DM) but still government and private sectors fear concerning individual privacy. Developing PPDM methods has become a demand of the time. A PPDM delivers singular privacy even though by permitting pulling out of valuable knowledge from data. Various methods can be used to permit PPDM. Among these methods, one such method is to alter the composed dataset before publishing, which provides the safeguard on individual records from being re-identified. Even if an opponent has advanced knowledge, he cannot be certain about the accuracy of a re-identification, when the dataset has been changed. Such class of privacy preserving method depends on the fact that the datasets used for data mining purposes do not really want to hold 100% precise data. Actually, that is certainly not the situation, because of the natural noise present in the data sets. In the context of data mining it is significant to preserve the patterns in the dataset. Furthermore, preservation of numerical factors, like covariances, means and variances of attributes is significant in the perspective of statistical data sets.

More data feature and privacy are twofold noteworthy desires that a decent

privacy preserving method needs to gratify. We have to estimate the degree of privacy and the data quality of a perturbed dataset. Data quality of a perturbed data set can be assessed over a little quality pointers like as extent to which the actual patterns are preserved, and preservation of statistical parameters. There is no sole agreed upon definition of privacy. So, defining privacy is a stimulating task.

1.1.2 Privacy Preserving in Data Stream Mining (PPDSM)

Data stream is a flow of boundless, real-time data substances with high data rate that can be merely read just once by means of an application. Streaming data are ordered in sequence and these data are endless. Such characteristics create issues in the area of data streams which are more challenging. Data Streams have different challenges as we know. Data streams need to be examined for recognizing trends and forms which support us in separating anomalies and forecasting upcoming behavior. Though, data proprietors cannot be ready to precisely disclose the real data values due to numerous explanations, most especially privacy considerations. Therefore, certain quantity of privacy preservation needs to be complete on the data before it can be made openly obtainable. During the process of data mining to preserve the privacy on data value, the problem of PPDM has been extensively studied and several methods have been suggested but, existing methods for PPDM are planned for static datasets and are not appropriate for streaming data. Hence the problem of privacy preserving of data stream mining is an essential issue. The goal of our thesis is to find solutions for privacy preserving clustering and classification on data stream with perturbation method and anonymize method. The proposed solutions should take into account the following:

- a. Various types of attributes, namely continuous, integer, nominal, and ordinal attributes.
- b. level of privacy
- c. time complexity of PPDM Techniques
- d. level of accuracy results

The main challenge is to decrease the accuracy loss with maximized the privacy level of the results caused by proposed heuristic based perturbation and anonymized methods.

1.2 Motivation

The extensive usage of information technology and online-medium has made it challenging for the individuals to gather, share and interchange data. Huge number of private data like shopping item details, criminal history, health history, credit details and others are extensively gathered and analyzed along with the improvement of data mining technologies. Such private dataset records are more important for different sectors whether it is private or government for taking crucial decision and providing societal prosperity like research in medical field, crime rate reduction and nationwide security. Mining the dataset records may be threat to our privacy, if we do not use it appropriately since data mining tools expose underlying pattern or all types of knowledge. The privacy anxiety has to turn into a key hurdle to information distribution. The consequences are dual: first, sectors like government and corporations worry about keeping extremely sensitive data private and therefore are not keen to share it with other sectors, not even declaring to publish to the community; second, different users grow cautious and distrustful of sectors that control sensitive data and therefore are not keen to submit their data to either sectors.

A large number of methods have been suggested for protecting personal privacy and sensitive information to overcome such an obstacle. Study of private data may be an attack on our personal privacy. With the growing of powerful data mining techniques and more and more information available on the internet, there are growing worries that DM (Data Mining) may pose damage to privacy and data security. As per our opinion, data mining tools should extract common patterns but should not reveal the private information of any person or sectors. In this wisdom, we trust that the actual privacy worries are with unrestricted access to individual records. Data mining tools that do include private data, in several cases, privacy is required to be preserved. Data mining can be respected in many applications, but due to no adequate guard data may be mistreated for further goals. The key factor

of privacy breaking in data mining is data abuse. In fact, if the data consists of serious and private characteristics and/or this method is mistreated, data mining can be harmful for individuals and sectors. Hence, it is essential to prevent not only revealing private information but also the critical facts.

1.3 Objectives

During the process of mining, preserving privacy on data has a trade-off between maximize privacy and minimize the information loss. So the prime objective of this research is to achieve best possible privacy level with not much to compromise on data utility. Proposed research work is carried out with the following objectives:

- a. To develop framework/algorithm to modify data that preserve privacy without sacrificing much on data utility.
- b. To use geometric data perturbation and sensitive drift approach for preserving privacy in data stream mining.
- c. To combine Sensitive Drift approach with multi-iterative k-Anonymization for preserving privacy in data stream mining.
- d. To develop algorithm which is applicable to various types of datasets and data mining techniques.

1.4 Scope of the Research Work

Following ways have been resorted to in order to handle the issue of preserving privacy while mining data:

- a. Perturb only sensitive attribute values and replace sensitive attribute values with the perturb one.
- b. Remove identifier attributes that directly reveal personal identity from dataset. Find out set of quasi-identifiers. Keep sensitive attribute values unchanged and anonymize all quasi-identifiers values.
- c. To use standard datasets followed widely by data mining research communities to test proposed algorithms.

- d. To evaluate performance of proposed algorithms on standard parameters.

1.5 Major contributions

The main goal of this thesis is to design and implementation of privacy preserving data stream mining techniques and to help various organizations to find ways to share and mine user data for the purpose of discovering patterns make better decisions while protecting their user's privacy, and to inspire as well as help organizations reason about the privacy-utility trade-off which maximize the privacy with minimum information loss. To achieve this we have proposed privacy preserving data stream mining algorithms, and analyzed the trade-off between utility and privacy using various privacy and utility measurements. The contributions of this thesis are:

- a. First contribution of this thesis is to design and implementation of heuristic based privacy preserving data stream mining using hybrid geometric data perturbation scheme. Proposed approach is achieved the privacy on sensitive data using translation, rotation and scaling with different orders. The goal of our proposed approach is to maximize the privacy with minimum information loss.
- b. Second contribution of this thesis is to design and implementation of SD - Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in data stream clustering scheme. Here, we introduced concept of sensitive drift (SD) based perturbation to achieve the privacy on sensitive values.
- c. Third contribution of this thesis is to design and implementation of heuristic based hybrid privacy preserving data stream mining approach using Perturbation and multi-iterative K-anonymization.
- d. Performance evaluation in a simulated large-scale deployment using MOA (Massive Online Analysis) tool.

The outcome of this work is published as papers in the International journals.

In this chapter, we have provided a brief introduction of Data Mining, Data Stream Mining and its application. Also, besides the privacy problems related to DM (Data Mining), Stream Mining and the growing public concern regarding their privacy are also discussed. Because of the vast public concern we want PPDM. We have also given a brief motivation, objectives of the proposed research work and scope of the work. In the next chapter, we will present a background study on privacy preserving data mining and data stream mining.

1.6 Organization of Thesis

The rest of the thesis is organized by Sequencing of the remaining chapters.

Chapter 2 covers the background knowledge of privacy preserving data mining and stream mining like, Data Mining, Data Stream Model, Data Stream Mining, Data Perturbation, Quasi-identifiers and K-Anonymity.

Chapter 3 elaborates study and observations which are made from research work carried out in the area of PPDM.

Chapter 4,5 and 6 discuss proposed algorithms to preserve privacy on streaming data by perturbing sensitive attribute values or by anonymizing quasi-identifiers without modifying sensitive attribute values. Algorithms have been tested with standard data sets and data streams. Accuracy and efficiency of proposed algorithms and their outcomes have been evaluated against standard parameters like Precision, Recall, Bias in Mean (BIM), Bias in Standard Deviation (BISD) and Confusion matrix.

Finally, in Chapter 7, Research work has been concluded with mention of future scope of expansion.

Chapter 2

Background Study

In this chapter, we present the brief introduction about data mining, data stream model, data stream mining and basic definitions of the key terms which will be used in the rest of the study.

2.1 Data Mining

Data mining alludes to the way toward separating novel, already obscure and conceivably valuable learning from the vast volume of data. It utilizes sophisticated methods for the way towards sorting with a lot of data sets and selecting pertinent information. Data mining applications help to make proactive, knowledge-driven decisions by predicting future trends and behaviors. Data mining finds out stimulating and unseen forms from enormous volumes of data from various sources like databanks, data repository, OLAP (on line analytical process) or further statistics depositories. With the emergence of sophisticated tools and technologies for data collection, the amount of data is increasing many folds every year. With such a huge available data in archives, extracting information for decision making using data mining is becoming a need for the time. Data mining has progressed because of the extensive growth of product and research work process. This evolution initiated when industry and other data was first deposited on systems and more newly, produced technologies that make available real-time steering of their data.

Few ways of defining the data mining are: 1) KDD (Knowledge Discovery in Databases) (Fayyad, Piatetsky-Shapiro, and Smyth). 2) DM (Data Mining) which contains methods from several disciplines such as statistics, database applications,

(ML) Machine Learning, (NN) Neural Networks, information retrieval, etc. (Han, Pei, and Kamber). 3) Data mining is the method of determining meaningful patterns and relationships that lies unseen within very huge databases (Seidman). 4) Data Mining is the examining of observational datasets to discover unimagined relationships and to recapitulate the data in novel ways once are each understandable and useful to the data proprietor (Hand, Mannila, and Smyth). There are many fields like, government, business sector, education, sports, share market, retail business, wireless communication (e.g. mobile, satellite) and transport where the data mining is widely used.

2.2 Data Stream Model

These days, we have seen various existing of sources of data delivered uninterruptedly at high speed. the instances are network activity, Global positioning system records, calling records (mobile), E-mail messages, sensor systems, client click streams, and so forth. These data sources are described by continuously producing gigantic measures of data from non-stationary distributions. Querying, repository and maintenance of streaming data got new challenges in data mining and database societies. Database groups have established DSMS (Data Stream Management Systems) for endless querying, sketches and summaries, sub-linear procedures for enormous dataset examination.

We assume that the input of data stream model comprises of several nonstop streaming data. Without loss of simplicity and analysis purpose, we may assume that every single tuple comprises of a solo attribute; the input consists of N data streams denoted as $D^1, D^2, D^3, D^4, \dots, D^N$ For any i^{th} data stream D^i , The stream gathering is written as $D = [D^i \text{ for } 1 \leq i \leq N]$. The stream gathering D can be measured as a $T \times N$ matrix where T is the present timestamp and N is the number of streams, which rises forever.

2.3 Data Stream Mining

The remarkable progress of the internet has produced a condition where vast amount of information is being swapped endlessly in the form of data streams. Collect the data at one place and after-process approach is often impracticable in given envi-

ronment due to massive data rate or real time applications. DSMS (Data Stream Management System) is being designed to support advanced applications on endless data streams which is a novel generation of information management system. Data mining methods are appropriate for straightforward and organized datasets like interactive databases, data repository and transactional databases. Due to rapid and non-stop advancement in database technologies, data repository applications and online medium, makes data produce speedily in various and complicated forms like none/semi structured, spatial/temporal, hypertext/multimedia. It is an important task to mine such complicated data in data mining field.

for last few years, many works are suggested (Babcock et al.) (Muthukrishnan et al.) to overcome the issues of treating and caching the data of endless and fast streams of data . Data Stream mining (DSM) states to informational framework pulling out as models and patterns from endless data streams. There are various challenges in many aspects of data streams like, storage, computational, querying and mining. Based on latest work in data stream mining, it is essential to plan novel methods to substitute the old ones because of data stream requirements. Traditional data mining techniques would demand the data to be cached and processed offline by means of mining procedures that create number of pass over the static data, however in data stream mining, streaming data is boundless and data-items produces with high speed, so it can be problematic to accumulate it. For that reason core challenges are dual (Aggarwal), (Chu) : First, many issue are caused because of vigorous nature of data streams, whereby the application of stream mining essential to identify varying concepts and data distribution and to adjust them. Second, Planning fast approaches for mining data streams and it is essential to regulate varying concepts and data distribution because of dynamic nature of data streams. For data streams, planning light and fast mining methods are most crucial issue; For instance, procedures that merely need single pass over the data and work with inadequate memory. In, traditional system of data mining, it typically needs whole dataset to be present, random access or multiple passes to the data, ample time per data item. On the other hand, there are few issues of DSM (data stream mining) that are impossible to store the entire data, random access is costly, simple calculation per data due to time and space constraints. The over-all

method of data stream mining is illustrated in figure 2.1.

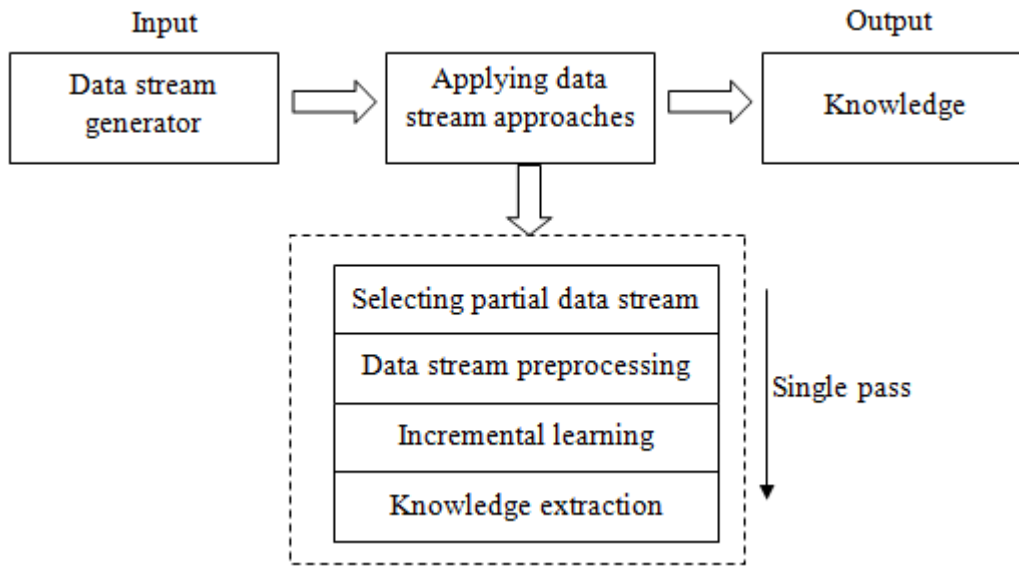


Figure 2.1: The general process of data stream mining

2.3.1 Data Stream Classification

It is necessary that the classification procedure must meet a number of requirements in order to work with the expectations and to be proper to learn from data streams. Following figure 2.2 shows the data stream classification cycle with the requirements (which is shown from 1 to 4). (Bifet et al.).

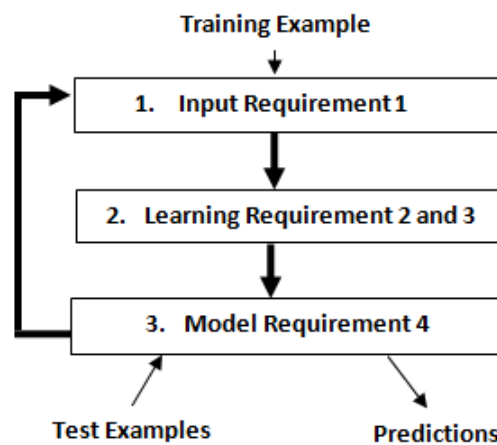


Figure 2.2: Round of Data Stream Classification

- Examine the instance only once (at most) and process it at a time.
- Work in a limited amount of time

- Workout with inadequate volume of memory
- Be prepared to forecast at several point

Framework of classification can be categorized into two: 1) Non-Incremental learning and 2) Incremental learning. In first, the training dataset are not entirely received at one time. Classification model is built via whatever data has been received, and revise the classification model based on recently received data. Incremental learning can adequate for data stream (Utgoff) while In second, unlike incremental learning, once entire data is totally deposited, few of them work as the training data to build a classification framework. It has greater computation cost and is not capable to gratify operator desires that want instant response.

For a number of latest applications, such as sensor based systems, electronic-mail, planning of schedule, intrusion detection and others., non-incremental learning is not suitable because of the incapability to gain whole training data before building the model of classification. The cost of model construction will increase enormously because it is essential to rebuild the classification model whenever fresh data is achieved. on the other hand, changing the model of classification to adapt fresh data is a more effective and feasible way. Incremental learning are classified into three ways.

- Learning without keeping instances is the first class of incremental learning (Schlimmer and Fisher). Each and every time fresh data is acquired, old data is abandoned. But, the classification model is not completely uncontrolled. As a substitute, fresh data is combined in the classification model. Shortcoming of the classification model will be unable to remember some formerly learned cases. In addition, the same training dataset may create diverse classification rules since the order of finding data is diverse.
- learning with partial instance memory is the second class of incremental learning. (Maloof and Michalski) presented the AQ-Partial Memory learning technique, which deposits data located to close the rule boundary. Fresh data are joined with deposited data as training data to change the classification model after coming data.

- learning with whole examples is the third class of incremental learning (Jin and Agrawal). In the period of the process of learning, whole streaming data is well-preserved and the streamin data which is used to decide whether the best attribute is reside in each node. Upon entry of stream data, fresh data are tested alongside old data. Modifying the test attribute via modification method, if the test attribute is not best attribute.

(Street and Kim) established a stream ensemble procedure for classification. The procedure separate data into a number of fix sized endless pieces in first step. Then, it builds a classification model for every single individual piece of data. Lastly, an ensemble classification model is assembled by merging a number of singular classification models. In (Domingos and Hulten) suggested a method called VFDT (Very Fast Decision Tree Learner). It practices the statistical outcomes of the Hoeffding bounds (Maron and Moore) to define with smaller amount of samples if the alteration among the outcome of the top attribute value and that of the second top test attribute is superior to a variance value. The key disadvantage of this approach is its incapability to handle data distribution from diverse time. The VFDT procedure cannot measure the time of data, and so, do not mine the data from changing time. In (Gama, Rocha, and Medas) proposed the method of VFDTc, which advances the VFDT method in two ways: the usage of an additional influential classification method and process endless values in the leaf nodes. This method still has some shortcomings like, in certain applications operators may only be interested in data that arrive in a certain period of time. In (Hulten, Spencer, and Domingos) suggested CVFDT method which improves the shortcoming of supposing data which are stably disseminated and also extends the features of the VFDT procedure.

2.3.2 Data Stream Clustering

Traditional Partitioned based method such as K-Mean, K-Mode, K-Medoid, Clarns etc, Hierarchical based method such as ROCK, BIRCH and others, Density based method like Optics, Dbscan, Denclue and others have been used in applications where tenacious data available and learning models creates are static in nature.. Traditional data clustering approaches are not efficient in the data stream model

since they are not capable of addressing the various issues (as discussed above) of data stream model. However refinement of few of above algorithms is also proposed in the field of data stream clustering model.

K-Mean (Ordonez) can be measured as primitive technique in this field. Actually K-Mean algorithm is proposed for traditional data mining. The same idea is also applied for data stream clustering. Randomly select the k-object as initial cluster. For each of the remaining objects, it allocates the object to the cluster, based on the distance between the object and mean of cluster and then estimates the new mean for all object with iterative manner. It has really an advantage of simple implementation, it also suffers from inherent limitations such as sensitive to outlier, not suitable for generating the cluster with arbitrary shape, not efficient for data for categorical attributes, selection of initial cluster mean is crucial etc. However various variants of K -Mean have seemed to address such obstacles such as Scalable K-Mean (Bradley, Fayyad, Reina, et al.), Online K-Mean (Sato and Ishii), Incremental K-Mean (Ester et al.) , (Ordonez) has suggested an improved incremental k-means method, HKA (Mahdavi and Abolhassani) etc.

(O'callaghan et al.) has proposed Stream methods for best quality data stream clustering for K-Median Problem as a Incremental K-Mean. It is a single pass algorithm. Stream algorithms process the data stream in buckets of m points it then precises the container information by holding merely the information concerning the k-center, with every single clusters center being weighted by the amount of points allocated to its cluster. Stream holds merely center information and removes the points. This is repetitive at every single level at utmost m points are reserved. Stream improves the K-Mean in the area of limited space and time with a constant factor $O(k_n)$. however Stream clustering algorithm considers neither time granularity nor evolution of the data. The clustering can become dominated by the older, outdated data of the stream. To address the issue of clustering evolving data stream of Stream (O'callaghan et al.), (Aggarwal et al.) proposed a new framework Called Clustream. Given algorithm separate the process of clustering into offline and online mechanisms. The online mechanisms work-out and store data stream summary statistics by means of micro-clusters and performs incremental online computation. The offline component does macro clustering answer the query us-

ing stored summary statistics which are based on the tiled frame model (Aggarwal et al.). A micro cluster in Clustream is represented as a clustering feature. Clustering features is a temporal extension of BIRCH (Zhang, Ramakrishnan, and Livny). When a new point is arrived, it may assign to existing cluster or a new one. If the new point fall within the boundary, it is added to the cluster. If not then the cluster is created. Two existing clusters have to merge or the least newly used existing cluster has to be removed. Macro clustering allows exploring the streaming clusters in excess of varying time skylines when the changes are dramatic. Clustream algorithm is scalable in terms of dimensionality of stream size and the amount of cluster. The tiled time edge model along using micro-clustering will result in improved efficiency and accuracy on real data. But the drawback of this algorithm is that, it generate the only spherical cluster.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchie) (Zhang, Ramakrishnan, and Livny) is a Hierarchical based clustering method to addresses two issues in clustering methods: Scalability and Incapability to undo what is done in the earlier step. This technique presents new concepts in two ways: macro and micro clustering. Its mechanism is based on two steps: first it scans database and makes a Cluster feature tree (Zhang, Ramakrishnan, and Livny) which is viewed as multilevel compression of the data. In second step BIRCH improves tree by eliminating sparse nodes as outliers and concrete real clusters. Clustering feature tree, Cluster feature vector make BIRCH effective for boosting and vigorous clustering of incoming data. Key drawback of this scheme is the constraint in volume of leaf. BIRCH does not perform well if clusters are not spherical in shape. Cobweb is originally proposed by (Fisher), but now refined and known as an incremental model for clustering using hierarchical conceptual ways. Hierarchical clustering model which is in the form of classification tree keeps by Cobweb. The method of Cobweb works using Category Utility function (CU) (Fisher) that measures clustering excellence. Outliers can be managed relatively well in this. Cobweb can be used to predict missing attributes or the class of a new object. Overhead of this algorithm is managing tree. A factor allowed stream clustering method is proposed in (Kranen et al.). Clustree is proficient of treating the stream in a single pass and continuously keeps an up to-date cluster model, concept drift, and outliers. Clustree is an

anytime clustering method. Another Micro-Macro data stream clustering method which is compactness based clustering is discussed in (Cao et al., “Density-based clustering over an evolving data stream with noise”). DenStream has a capability of dynamic adaptability to change of clusters, limited memory usage, cluster with arbitrary shape, detecting and separating outliers. The Algorithm is divided into: (1) Online micro-cluster maintenance, (2) offline Micro-cluster Maintenance. DenStream introduced o-micro-clusters to maintain distinct memory space, which is called an buffer-outlier. Density based clustering algorithm (D-Stream) which is presented in (Tu and Chen) can determine randomly formed clusters; algorithm can control noises and it is a single scan method.

2.4 Data Perturbation

Privacy preserving on customer’s data during data mining process is done in two ways: perturbation and anonymization. Perturbation means modify the sensitive values using various proposed methods. Such scheme like random noise data is announced to change sensitive data values and the spreading of the random data is used to produce a new data sharing which is near to the real data distribution without disclose the real data values. Statistical outcome is mostly the same during the process on perturb and real data values means, the statistical information-gain designed from the perturbed data, does not vary from the statistical information-gain designed from the real data. The perturbed data records do not match up to real-world record proprietors, so the opponent cannot accomplish the sensitive links or recover sensitive information from the existing data.

2.5 Quasi-Identifiers (QI)

A Data proprietor can frequently recognize attributes in their data that also seem in external sources, and such attributes are nominees for connecting which is entitled as Quasi-Identifiers (QI) and it is basically the mixtures of such quasi-identifiers that must be protected. quasi-identifiers (QI) which could be linked with external source to re-identify individual record owners.

2.6 K-Anonymity

One of the important types of privacy outbreaks is re-identification of individuals records by means of quasi-identifiers (Samarati and Sweeney). This type of attack is solved by anonymization. K-anonymization is one of the methods which is suggested by (Samarati and Sweeney), (Sweeney). The idea behind k-anonymity is to suppress or generalize the selected data which are publicly available so that each of the records becomes very similar from at least $k - 1$ other record. Consequently, the sensitive data may be connected to sets of records of size at least k . We consider tabular data where each row is an individual, and the columns (attributes) are labeled "sensitive" or "insensitive". We want to protect the sensitive attributes. For example, in the table 2.1, "disease" is a sensitive attribute and the others are insensitive. Quasi-identifier attribute values are a minimal data attribute value set that linked with other data set can uniquely identifier individual. K-anonymity is to prevent individual privacy without changing sensitive attribute values. The traditional k-anonymity primarily for static data set and cannot be applied to data stream directly. In our proposed work, we revised the traditional k-anonymity definition to fit for data stream. Another data stream concept about k-anonymity is anonymous data delay, which primarily consider the maximum time of tuple keep on in memory. Table 2.2 is an example of 2-anonymity applied on the left one. In general, the following rules are applied:

- Rows are clustered (partitioned) into sets of size at least k
- Within each set, make insensitive attributes identical. There are usually two ways of doing so: 1. Suppression: delete an entry (e.g., let Gender attribute be null). 2. Generalization: replace with less specific info (e.g., for Age, substitute [40,49] for 42).
- Sensitive attributes remain untouched.

Table 2.1: Medical dataset

Gender	Age	Zip	Medical Condition
F	42	13155	Hepatitis
M	45	13144	Diabetes
M	33	12346	Heart
M	30	12345	Heart

Table 2.2: Anonymized dataset (2-Anonymity)

Gender	Age	Zip	Medical Condition
*	4*	131**	Hepatitis
*	4*	131**	Diabetes
M	3*	1234*	Heart
M	3*	1234*	Heart

2.7 MOA-Massive Online Analysis

MOA provides a software environment that helps to implement methods and running tests for online learning through classification and / or clustering from data streams (Bifet and Kirkby)(Bifet et al.). MOA is intended to handle the stimulating issue of scaling up the implementation of state of the art procedures to real world data set dimensions. It encompasses many variants of classification and clustering for data streams mining and tools for evaluation as well. Researchers can benefit from MOA by getting in-depth knowledge about workings and problems of different approaches. Users can, without much of a stretch apply and match numerous procedures to real world dataset and factor settings.

2.7.1 Characteristics

- MOA is an open-source application and framework for practice, hands on research. MOA has bidirectional interface with WEKA-tool.
- It contains number of algorithms which are implemented for testing and comparison to approaches from the literatures.

- It provides an interface with standard streaming data sets via warehoused, distributed settings for the various data inputs and noise options, both real and synthetic.
- MOA is developed in Java programming language and the advantage of Java is portability.

A data stream environment has altogether various needs from the traditional batch learning. The main noteworthy are:

- Real time processing, and at the most one inspection
- Use of a inadequate quantity of memory
- Less execution time
- All the time ready to predict

2.7.2 MOA-Features

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

2.7.2.1 Data Stream Generators

Generators generate the MOA Streams, MOA streams can be available .ARFF files, merging the numerous data streams, or purifying set of streams. Available data stream generators are: *Random Tree*, *SEA Concepts*, *STAGGER Concepts*, *Rotating Hyper-plane*, *Random RBE*, *LED*, *Waveform and Function Generator*

2.7.2.2 Classifiers

The implemented classifier methods includes- *Bagging using ADWIN and Bagging using Adaptive-Size Hoeffding Trees*, *Nave Bayes*, *Decision Stump*, *Hoeffding Tree*, *Hoeffding Option Tree* and *Bagging and Boosting*.

2.7.2.3 Clustering

MOA encompasses an investigational framework for clustering streaming data, which permits matching diverse methods on separate data stream settings or comparing same data streams settings on different approaches. Major stream clustering features of MOA are:

- Data generators for growing data streams
- extensible number of stream clustering procedures
- Assessment measures for stream clustering
- Visualization tools for analyzing outcomes and matching different parameter settings

MOA encompasses number of clustering techniques like, *CobWeb*, *CluStream*, *DenStream*, *D-Stream*, *StreamKM++*, *ClusTree*. Moreover, MOA encompasses evaluation measures for analyzing the performance of the generated clustering models. The available measures evaluate the accuracy of the resulting clustering apart from this it also verifies the correct assignment of examples.

The component of visualization in MOA, visualizing the stream and clustering results. Figure 2.3 and 2.4 provides performance comparison of two different clustering algorithm on the same stream settings with six online evaluation measures- F1_P, F1_R, Purity, Precision, Recall and Redundancy. The above part of the Graphical User Interface (GUI) where, pause and resume button for intermediate evaluation of streams, speed adjustment, Dimensions selection and components to be displayed like points, micro-and macro clustering and ground truth to be displayed. The below part of the GUI shows the measured values for both settings as values (left side, arithmetic mean values) and the recently selected measure as a plot over the arrived set of instances as examples (right side, purity measure in this example).

2.7.3 Evaluation Measures

MOA provides several cluster evaluation measures (see 2.4). Precision/Recall are normally used to regulate the usefulness of the system. Method returned considerably more relevant outcomes than irrelevant if precision is high and method returned most of the relevant results if recall is high. The greater the precision and recall, system tends to be more effective. Algorithm 1 and algorithm 2 for PPDSM have been evaluated on these measures.

$$Precision(f1_p) = \frac{(\sum_{I=1}^{|C|} f1_p(I))}{Realclust}(1)$$

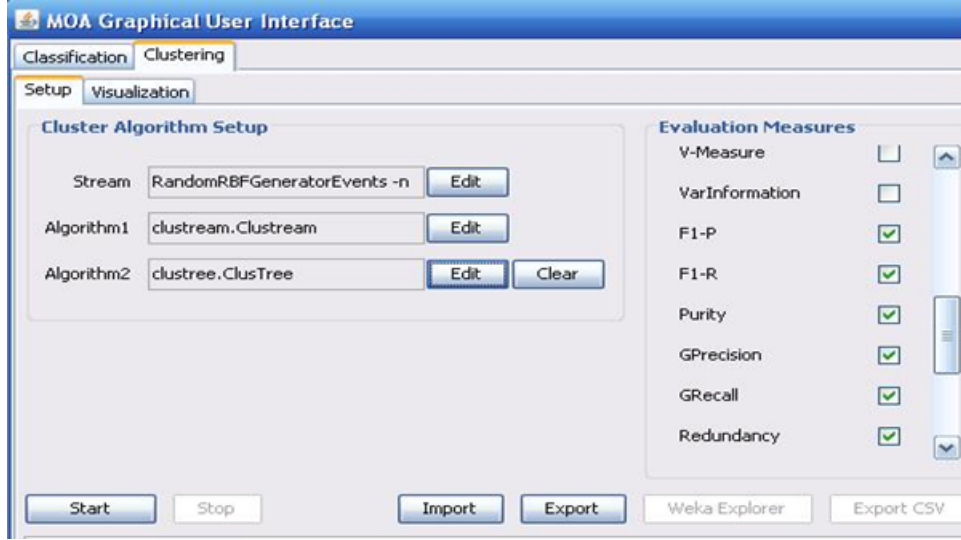


Figure 2.3: MOA data stream clustering setup

Where,

$|C|$ = Number of clusters

$$f1_p(I) = \frac{2 \times Precision(I) \times Recall(I)}{Precision(I) + Recall(I)}$$

$$Precision(I) = \frac{Max(C[I, J])}{Clustertotalweight(I)}$$

$$Recall(I) = \frac{Max(C[I, J])}{Classtotalweight(I)}$$

$C[I, J]$ = Two Dimensional array Represent cluster i, class j. return the weight of cluster I and class J.

$Clustertotalweight(I)$ = Returns the total number of instances belong to cluster I.

$Classtotalweight(I)$ = Returns the total number of instances belong to class J.

$$Recall(f1_r) = \frac{(\sum_{j=1}^{|C|} f1_r(J))}{Numclasses}(2)$$

Where,

$|C|$ = Number of clusters

$$f1_r(J) = Max(f(I=1,2,3,...Number\ of\ Clusters, J))$$

$$f(I, J) = \frac{2 \times Precision(I) \times Recall(I)}{Precision(I) + Recall(I)}$$

$$Precision(I) = \frac{Clusterclassweight(I, J)}{Clustertotalweight(I)}$$

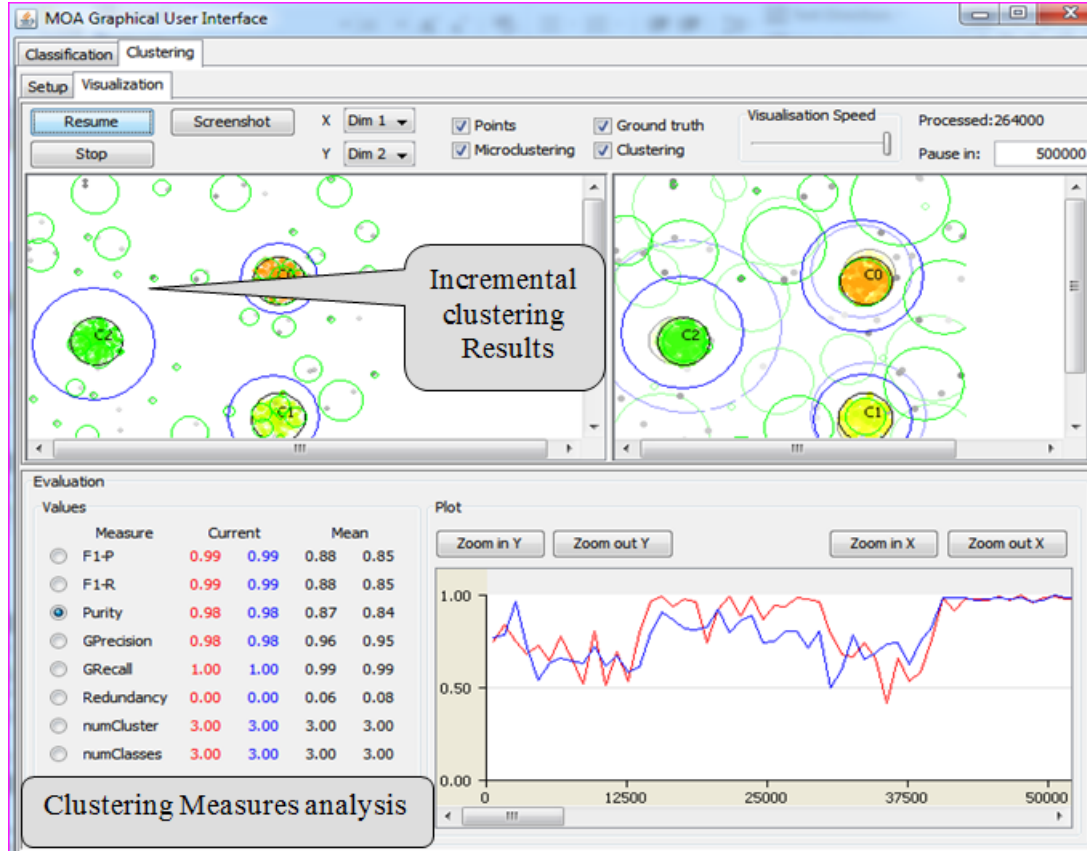


Figure 2.4: MOA visualization components

$$Recall(I) = \frac{Clusterclassweight(I,J)}{Classotalweight(I)}$$

$Clusterclassweight[I, J]$ = The total instances belong to cluster I and class J.

In this chapter, we have discussed various terminology as background study like, Data Mining, Data Stream Mining, Data stream clustering, Privacy Preserving Data Mining (PPDM) and Privacy Preserving in Data Stream Mining (PPDSM), K-Anonymity, MOA Framework etc. In the next Chapter, we will discuss about related work in the area of PPDM and PPDSM.

Chapter 3

Related work

In previous chapter, we have discussed different terminologies as background study. In this chapter, we have briefly described existing work related to privacy preserving in data mining and data stream mining research and we have summarized the contributions made by various researchers. Finally, we included research issues and challenges in the area of privacy preserving in data mining.

Typically when the person talks about privacy of data, one means that keep data or information should be kept secret from others. The opponent always use private or individual data that harmfully affects somebody's life. Most of the people do not longer want that their privacy should be dishonored. The issue crops up once information is published and it is not possible to protect abuse. Using this distinction, confirming that a data mining scheme would not allow misuse of individual information, unbolts chances that whole privacy would prevent. We need social and technical explanations that data will not be published. The same basic worry also applies to collections of data. Learning from data repository should not expose important information of individuals. One may not care about somebody knowing their SSN (Social Security Number), PIN code, birth date, Father or mother's initial name; but knowing all of them allows individuality burglary. This kind of privacy issues arise with huge, multiple singular collections as well. A method, that guarantees that, no individual data is discovered. Commercial data is normally the aim of data mining, but certain outcomes may still lead to concerns related to privacy. If we outlook expose of sensitive information about an individual as a likely individual privacy destruction, then generalizing this to

expose of information about a subset of the data captures both views. Security and Privacy protection have been a public policy concern for years. In 2000, the author presented two paper on Privacy Preserving Data Mining. Both papers addressed similar issues. The first paper was based on altering the data values so actual values are not revealed (Agrawal and Srikant) . The second paper was based on Secure Multiparty Computation to encrypt data values (Lindell and Pinkas) safeguarding that no party can pick up anything about another's data values. The objective of PPDM methods is to remove related information from data set while guarding sensitive information. Recent work in the field of PPDM has made an effort to achieve precise privacy at some level and maximize the information gain. Numerous efforts have been made to incorporate privacy preserving approaches with data mining processes in order to prevent the expose of sensitive information. PPDM techniques undoubtedly depend on the privacy definition, which identify what data is sensitive in the real data set and protects that data from either direct or indirect exposure. PPDM should enforce: 1] A PPDM procedure should have to elude the discovery of sensible statistics. 2] It should be robust to the numerous data mining approaches. 3] It should not settle the access and the use of non-sensitive data. 4] It should not have an exponential computational difficulty. Data perturbation means data alteration procedure and normally accomplished by the data proprietors on previously issuing data. The objective of executing such data alteration is twofold. Firstly, the data proprietors need to modify the data in some manner to cover the sensitive data in datasets, and secondly, the data proprietors need to alter it to best preserve those domain specific data properties that are critical for constructing expressive data mining models.

In the field of PPDM, perturbation methods are welknown method for privacy preserving on data or data stream. It is particularly beneficial for applications where data proprietors want to contribute in cooperative mining and at the same time need to prevent the outflow of privacy sensitive information in issued data sets. A wide varity of perturbation methods have been suggested (Agrawal and Srikant), (Agrawal and Aggarwal), (Chen and Liu, "A random rotation perturbation approach to privacy preserving data classification"), (Evfimievski et al.), (Feigenbaum et al.), (Lindell and Pinkas), (Vaidya and Clifton, "Privacy-preserving

k-means clustering over vertically partitioned data"). Among these methods, the wellknown method is the randomization method that emphasizes sole dimensional perturbation and assumes no dependency between data columns (Agrawal and Srikant), (Evfimievski et al.). In recent times, the data management group has displayed minor improvement on multiple dimensional data perturbation methods, like the condensation approach using KNN (k-nearest neighbor) method (Aggarwal and Philip), kd tree based on the multi-dimensional K-anonymization (LeFevre, DeWitt, and Ramakrishnan) and perturbation using the multiplicative data (Oliveira and Zaïane), (Chen and Liu, "A random rotation perturbation approach to privacy preserving data classification"), (Liu, Kargupta, and Ryan), (Chen, Sun, and Liu). In single attribute based data perturbation method that assume data attributes to be independent while, In multi-dimensional data perturbation method that assume data attributes to be inter-attribute dependency and distribution. Privacy preservation techniques are classified into five different dimensions (Verykios et al.).

- a. *Privacy preservation*: It is the most important method which is used for the watchful reconsideration of the data. The methods are: heuristic based methods such as adaptive alteration that alters nominated values that increase the information gain rather than existing data values. SMC (Secure Multiparty Computation) is a cryptography based method in which groups do not know except its own contribution and the outcomes at the end of computation, and reconstruction based methods where the actual distribution of the data is reconstructed from the randomized data.
- b. *Data mining algorithms*: The best essential thoughts have established data mining procedure which is based on classification, induction with decision tree, mining procedures using association rule and clustering procedures.
- c. *Data Rule hiding*: It mentions whether an input or outcome as aggregated should be hidden. The complexity for hiding aggregated data is greater.
- d. *Data modification*: Alter the original values of a data set before releasing it to public domain and in this manner safeguard privacy protection. It includes methods like; Perturbation, Blocking, Aggregation, Swapping and sampling.

- e. Data distribution: There are two types of methods. The first one is used for centralized data while the other type is used for scattered data scenario. Scattered data scenario can be categorized into parallel data distribution and perpendicular data distribution.

Multiplication or Noise addition is not the only method which is used to perturb the data value. There are other available methods also available like data swapping, in which the values are exchanged to accomplish the privacy preservation (Fienberg and McIntyre). One benefit of this method is that the basic statistical characteristics of the data are fully preserved and are not altered at whole. So, some kind of aggregate executions can be accurately run without compromising the data privacy.

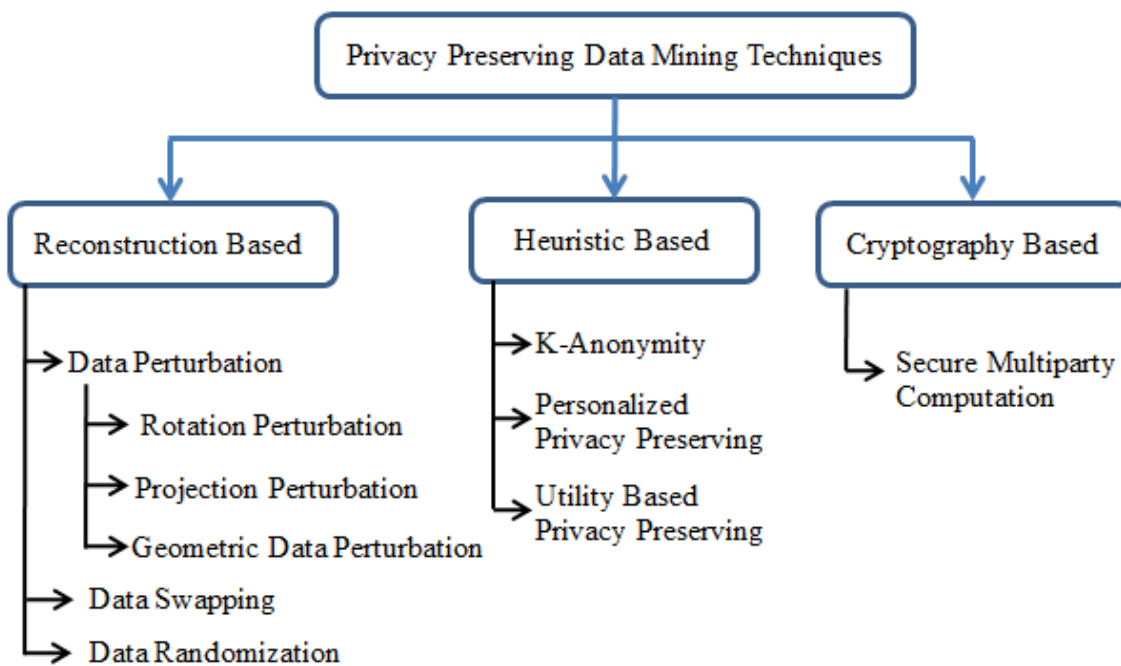


Figure 3.1: Privacy Preserving Data Mining Techniques

There is not an informal way to outline privacy and can be preserved on various levels in many scenarios in spite of enormous diversity in privacy facets of data mining. Three methodologies can be distinguished: 1) Heuristic based 2) Reconstruction based and 3) Cryptography based (see figure 3.1). The heuristic method is considered for centralized data. The cryptography based method is used for the distributed data, while the reconstruction based method can be applied to both dis-

tributed and centralized data. Figure 3.2 shows the classification of PPDM based on Data Hiding and Rule Hiding.

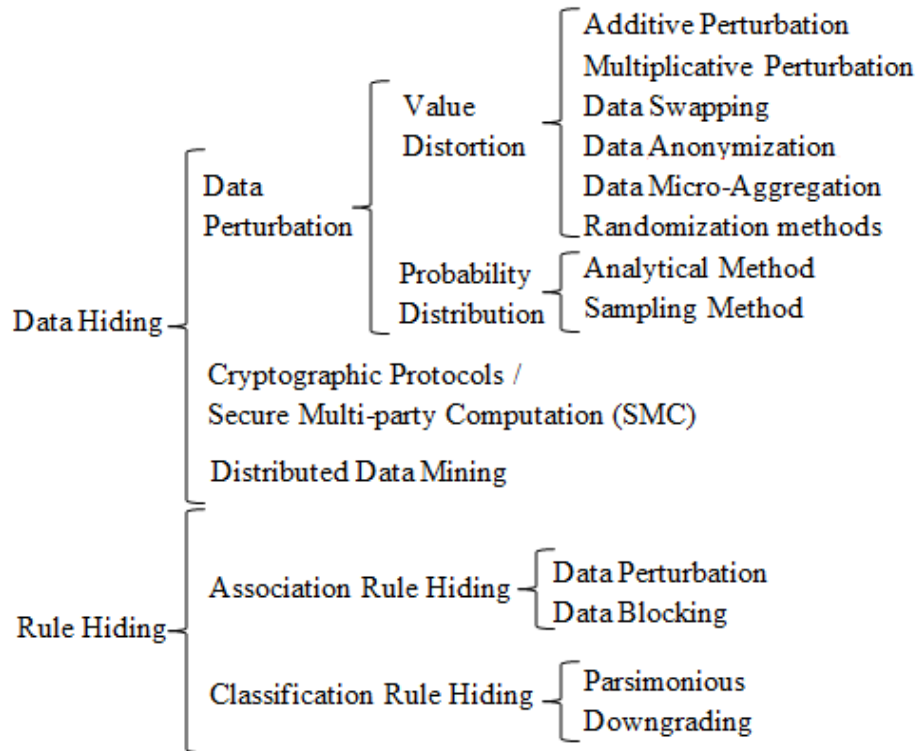


Figure 3.2: Data Hiding & Rule Hiding based PPDM Techniques

3.1 Heuristic Based Methods

In heuristic based method, the heuristic processes are used to hide sensitive data or information. Any organization does not want to disclose their crucial data. Individual values in data are altered according to a heuristic process to hide sensitive data or knowledge. In database community, researchers produced different methods that processed records in a "cluster based" manner, using information about particular local records universally to convert the records in a manner which preserves particular privacy metrics. These altered records can then be distributed without fear of any kind of security attacks or breaches. There is a supposition that certain fields of a record encompass quasi-identifier (QI) attributes that distinctively recognize an individual related with the record, as well as sensitive attributes that must not be connected to the individual by an untrusted third party. One of the grouping based approaches is "k-Anonymity". Another two

approaches *l*-Diversity and "*t*-Closeness" are also recommended that aim on accomplishing the final state where "*k*" records appear exactly the same.

3.1.1 k-Anonymity

(Samarati and Sweeney),(Sweeney) proposed first K-Anonymity model which accomplishes privacy by imposing the restraint that each row of the released datasets should be identical from "*k*" other rows in respect of a certain set of attributes or records which is called Quasi-Identifiers (QI). This is typically accomplished by eliminating all or part of a field value (called suppressing), using some pre-specified hierarchy of values (called generalization) and inter-changing (swapping) values of some of the entries in the dataset. K - Anonymity defends in contrast to identity expose; it does not make available appropriate protection in contrast to attribute expose. General method of K-Anonymity suffered due to back-ground knowledge attack and homogeneity attack. So, author (Machanavajjhala et al.), proposed a variant of K-Anonymity which is known as *l*-diversity. It promises that, privacy in certain conditions where *k*-anonymity does not achieve like, once there's slight variety in the sensitive attributes or once the opponent has some back-ground information.

Proposed method seems to be prejudiced towards privacy at the cost of usability. Additionally, *l*-diversity considers complete values of a specified attribute in a related manner without going in detail for data distribution. This is barely ever occurring for actual datasets, the reason being that data values of attribute may be amply twisted. This may perhaps make it further problematic to produce reasonable -diverse demonstrations. Frequently, an opponent may utilize back-ground awareness of the worldwide spreading to create inferences related to sensitive data values in the datasets. Moreover, not whole attribute values are likewise sensitive. Suppose in pertucular instance, an attribute matching to a medical-condition might be further sensitive when the data value is constructive, as opposed to when data value is destructive. (Li, Li, and Venkatasubramanian) shows that Spreading of a sensitive attribute in every sameness group is close to the spreading of the attribute inside the whole dataset using *t*-Closeness framework.

3.1.2 Personalized privacy preservation

Private and public sectors may have vastly diverse limitations on the privacy of its records compared to a specific person, So, records in a known dataset are preserved in a different way for anonymization purposes. This implies that in anonymization, the value of k may be different with the particular instance. In (Aggarwal and Yu) has proposed a condensation based method for PPDM in the existence of parameter restraints on the privacy of the datasets. Given method builds clusters of miscellaneous size from the value of data, with the end goal that, every single raw-data lies in a gathering whose mass is in any event equivalent to its anonymity level. Consequently, unnatural-data are produced from every single cluster so as to generate a mock dataset with the similar collective distribution as the actual data. Author in (Xiao and Tao) proposed one more exciting framework of personalized anonymity in which somebody can identify the position of privacy for their sensitive data values. Given method accepts that a singular can identify a point of the DGH (Domain Generalization Hierarchy) as opposed to choosing the level of k -anonymity which can operate with.

3.1.3 Privacy preservation based on utility

The loss of knowledge may also be measured in terms of effectiveness (Utility). In (Kifer and Gehrke) first studied the difficulty of PPDM based on utility. The idea was to extend the cursed dimensionality by disjointedly distributing negligible matrices encompassing attributes which have effectiveness (Utility), however it is furthermore challenging for preservation of privacy purposes. The generalizations can be executed on the negligible matrices and the original matrices need to be equivalent. In (Xu et al.) proposed the method of data mining using local recording which is based on utility. The method is based on the fact that from software point of view, various attributes have varied utility. Closely all anonymization approaches are universal, where specific instance data is plotted to the equal generalized data. The data space is detached into many areas in local recoding and the plotting of the instance to generalize value is local to that area. Another alternative way to anonymize the data using utility based PPDM is that it residues beneficial for specific types of knowledge discovery process. For instance, author

in (Fung, Wang, and Yu) proposed a method for k-anonymization through means of loss of information metric as the measurement of utility.

3.2 Cryptographic Based Methods

There are two key reasons for popularity of cryptographic based methods. First of all, cryptography offers a well-defined structure for privacy, which consists of practices for demonstrating and quantifying it. Secondly, to implement privacy preserving data mining procedures, there exists a huge tool-set of cryptographic algorithms. Though, cryptography does not guard the outcomes of a computation, as an alternative, privacy outflows are prevented in the process of computation. The data record sources may be scattered across the network and therefore bringing them together at a centralized place may not be possible due to limitations in computational and communication resources. There are two distributed applications models which store data. One is perpendicularly partitioned data model and the other is straight partitioned data model. Distributed data mining provides methods to perform data mining in a distributed setting without bringing together the data into one location where privacy thoughts can disallow administrations from sharing their own data with any other party. Preserving privacy in distributed procedures arose as a solution to this problem by permitting parties to work together in pulling out of knowledge without any of the collaborating parties to have expose its own specific data items to any other party.

In (Goldreich) proposed techniques from SMC (secure multiparty computation) to privacy preserving in distributed data mining. (Yao) introduced first secure multiparty computation for secure circuit evaluation, in theory, to compute any function over data partitioned between two parties, without revealing anything to either party beyond the computed output. Though, data mining typically encompasses huge number of data substances, the communication costs of these protocols render them unpractical for these purposes. This has led to the search for difficult exact protocols that have effective communication complexity. (Lindell and Pinkas) use cryptographic techniques in their protocol. Their work differs from general secure multiparty computation in the sense that most computation is done locally by the individual parties. The protocol involves a small

number of secure evaluations of small-sized circuits (thus resulting in a protocol of low communication complexity). (Du and Zhan) presented a privacy preserving protocol for constructing decision trees for a two-party vertically partitioned database using secure scalar product. This was prolonged to the multi-party case by (Vaidya et al.). They also presented an association rule based privacy preserving mining protocol for perpendicularly partitioned data and Naive Bayes based privacy preserving classifier protocol for perpendicularly partitioned data (Vaidya and Clifton, "Privacy preserving naive bayes classifier for vertically partitioned data"). In (Kantarcioglu, Vaidya, and Clifton) presented a Naive Bayes classifier protocol for horizontally partitioned data. (Kantarcioglu and Clifton) gave an association rule privacy preserving mining algorithm for horizontally partitioned data and (Vaidya and Clifton, "Privacy preserving association rule mining in vertically partitioned data") proposed the same for vertically partitioned data. (Agrawal, Evfimievski, and Srikant) proposed a method for computing union, set intersection and equijoins for two parties. This method should be carefully chosen by keeping complexity of algorithms and communication cost into mind.

3.3 Privacy Preserving Data Stream Mining (PPDSM)

Data streams are often produced by many real-time applications, telecommunication networks, internet traffic flows, online banking and financial transactions, retail market, factory production process data, sensor based application data flows, satellite data, research lab data, electric power grids, engineering data and other number of dynamic environments. Data streams are tremendous and possibly infinite in volumes. These data-streams need to be analyzed for recognizing trends and patterns, which benefit us in isolating anomalies and forecasting upcoming behavior. Though, data proprietors or originators may not be willing to precisely uncover the genuine values of their data because of some reasons, most particularly privacy considerations. Therefore, some amount of privacy preservation needs to be done on the data before it can be made widely accessible. The understanding of data is significant and it is conjoined with the need to preserve privacy using appropriate algorithms. Various approaches have been suggested for this purpose like data perturbation, k-anonymity, association rule mining, masking and encryp-

tion. Existing techniques cannot be applied directly on data streams. Furthermore, in data streams applications, there is a need to offer robust assurances on maximum allowed interval between incoming data and its anonymous output with minimum data losses and maximum privacy gain.

Some researchers have worked on this scenario and have resolved the issue of data stream mining privacy protection for last 10 years (Chao, Chen, and Sun), (Chao, Sun, and Chen), (Zhang et al.), (Zhou et al.), (Cao et al., "CASTLE: A delay - constrained scheme for k s-anonymizing data streams"), (Li, Ooi, and Wang), (Poovammal and Ponnaivaikko). In (Cao et al., "CASTLE: A delay - constrained scheme for k s-anonymizing data streams") proposed the method CASTLE and mainly targeted at the supreme adequate intervals between the incoming data and the anonymous outgoing flows. Proposed technique builds update with confirming unidentified data on the extent allowed by the interval base on unidentified clustering of data stream, in that way progresses usefulness of privacy protection. In (Li, Ooi, and Wang) proposed the method called SKY which accomplishes the motive of protection of private-ness with the way of k-anonymization on streaming data. The proposed method presents a constraint factor "d" to limit the extreme deviation of open data of every single tuple, in a way that it finalizes privacy protection in good way, even though run with superior assurance for the legitimacy of the data. In (Zhang et al.) proposed KIDS model which is based on K-Anonymization technique to preserve the privacy of streaming data. Sliding window mechanism has been used to accumulate data that fall within. The proposed model uses distributed density of data to forecast forthcoming value of data, which in turn increases the precision of data and increases the usability. Characteristics of data streams is high arrival speed, endless data, sliding window continuously updating. Therefore, algorithm needs extra time to operate k-anonymity on sliding window directly. TDS (Top Down Specialization) hierarchy data structure has been implemented that requires updating TDS-tree for latest incoming tuple and not whole tuples in sliding window. (Chao, Chen, and Sun) proposed PCDS (Privacy Preserving Classification of Data Stream). Generally, method of the PCDS technique for privacy preserving streaming data classification is separated into two phases, one is pre-processing steps on streaming data and second is min-

ing process on streaming data. First phase objective is controlled by Data Streams Preprocessing System (DSPS) which is perturbing data streams. Second stage is controlled by the Online Data Mining System (ODMS) which work with WASW (Weighted Average Slide Window) method and make classification model through mine the perturbed data streams. Author also shows ASD (Average Squared Distance) and DBRL (Distance Based Record Linkage) Security Measurement which provides better outcome than other methods.

3.4 Summary of Major Research Contributions

Table 3.1: Heuristic-based techniques

Author(s)	Year	Approach	Findings
Samarati	2001	First proposed k-attribute anonymity approach for micro data release.	Data integrity has been provided through generalization and suppression techniques. Degrades data usability.
Domingo-Ferrer and Mateo-Sanz	2002	Approach for aggregate data release instead of micro data was introduced to prevent identity disclosure.	Information loss is major problem due to micro aggregation release. Limited to numeric data.
Bayardo and Agrawal	2005	k-anonymity is an NP-hard problem hence proposed optimal k-anonymity approach.	It does not address classification requirements.
LeFevre, DeWitt, and Ramakrishnan	2006	Proposed multidimensional k-anonymity model.	More efficient than Bayardo and Agrawal
Nabar et al.	2006	Query auditing method was used with query monitoring and denies if query processing compromising privacy.	Rather than sensitive data, sensitive rule has been protected.

Machanavajjhala et al.	2007	Authors proposed attribute level anonymity called ℓ -diversity.	Linkage attack is possible with k-anonymity as it creates clusters that leak information due to deficiency of diversity in the sensitive attribute. Biased towards privacy at the cost of data usability.
Wong et al.	2006	Author proposed the (α, k) -anonymity model for privacy preserving data broadcasting.	(α, k) -anonymity protects both identifications and affiliations to sensitive information in data.
Li, Li, and Venkatasubramanian	2007	Extension to k-anonymity and ℓ -diversity was proposed.	Spreading of a sensitive attribute in any proportionality class is close to the spreading of the attribute in the general table. Better privacy and data usability compare to k-anonymity and ℓ -diversity.
Poovammal and Ponnavaikko	2009	Surveyed anonymization techniques.	Methods provide secrecy at the cost of data usability.
Fung, Wang, and Philip	2007	k-anonymity model was used with decision tree.	Authors claimed better privacy with proposed approach.
Kadampur et al.	2010	Noise addition in decision tree classification was proposed.	Different set of algorithms for noise addition were presented for numerical and categorical dataset. Data quality in modified data is a big concern.

Zhong, Yang, and Wright	2005	Combined heuristic-based and cryptography-based approaches were used on distributed customer data.	Proposed protocol claims to provide end-to-end privacy.
Blum et al.	2005	Query audition framework was proposed where trusted administrator prevents access to private information by adding noise to the query response.	Approach does not perturb input data properly.
Ponnaivaikko and Poovammal	2009	Discussed methodology to protect sensitive value from heterogeneous medical Data set.	Focused on sensitive data protection. Proposed method can be applied to numerical and categorical Data set. Privacy is a big concern in certain conditions.
Zhang et al.	2010	Proposed framework for privacy mining using data stream k-anonymity. Sliding window mechanism is used.	Very less work has been done for data stream privacy.

Table 3.2: Reconstruction-based techniques

Author(s)	Year	Approach	Findings
Agrawal and Srikant	2000	Proposed reconstruction based approach via adding random values from a probability distribution for privacy of sensitive data.	Reconstruction from perturbed data set is possible through PCA and spectral filtering. Does not perturb categorical data.

Agrawal and Aggarwal	2001	Proposed reconstruction based algorithm using an EM (Expectation Maximization)	Given technique does not consider the dissemination of the original data.
Dutta et al.	2003	Data distortion through randomly via adding noise into original dataset.	Increase data utility in certain cases. Works fine if relative amount of noise is smaller.
Kargupta et al.	2003	Proposed single-attribute value random matrix to disturb value, through producing a multiple attribute combined distribution matrix to rebuild the original data set.	The real data set converted into a pseudo data set which can damage the secrecy.
Rizvi and Haritsa	2002	Proposed a framework called MASK for sensitive rule hiding.	Tradeoff between data utility vs level of privacy is an issue. Applied to association rule mining only.
Wu	2005	Randomization for specialized application where users do number of computation.	Tradeoff between data utility vs. level of privacy is an issue.
Agrawal and Haritsa	2005	Proposed FRAPP framework for association rule mining.	Tradeoff between data usability vs. level of privacy is improved compare to all earlier approaches.

Zhang and Bi	2010	r-amplifying and matrix condition number has been used to protect data privacy.	Applicable to centralized as well as distributed data distribution with numerical and categorical data types.
Guo and Wu	2009	Proposed randomized techniques for privacy preserving with unknown distortion parameters.	Works on categorical data only and extension of it on numerical data is still topic of research.
Mishra and Sandler	2006	Privacy via pseudo random sketches instead of entire perturbed data sample has been used for mining to preserve data privacy.	Focused only on AND queries.
Kamakshi and Babu, "Preserving the privacy and sharing the data using classification on perturbed data"	2010	Gaussian distribution along with noise addition is used to achieve better tradeoff between privacy gain and information loss.	Applied to classification data mining method with numerical data only.
Karmakar and Bhattacharyya	2009	Used both randomization and data perturbation techniques to modify data.	Different privacy levels can be adapted for different attributes. Works well only with centralized data. Biased towards privacy at the cost of data utility.

Kamakshi and Babu, "Automatic detection of sensitive attribute in PPDM"	2012	Present novel idea to randomly identify the sensitive attributes of PPDM. Recognition of sensitive attributes is subject to threshold limit of sensitivity of each characteristic.	The data is altered in such a way that the real properties of the data remain unchanged.
Zhang, Yang, and Chen	2012	Proposed HPNGS which is a fresh improved past probability based noise generation method.	Method is accomplished in decreasing noise requirements over its random complement.

Table 3.3: Cryptography-based techniques

Author(s)	Year	Approach	Findings
Lindell and Pinkas	2002	First proposed SMC protocol for data mining classification technique.	Worked for decision tree classification ID3 algorithm with privacy preservation. Focussed on securing two party protocol, with ID3. Communication cost is big overhead.
Vaidya and Clifton, "Privacy-preserving k-means clustering over vertically partitioned data"	2003	Vertical partitioning of data has been proposed with each site learning the cluster of restricted entities.	Used SMC approach with multiple users with K-Mean clustering approach. Faces the tradeoff between communication cost and level of privacy.

(Zhan) and (Ponnavaikko and Poovammal)	2008, 2009	Follow SMC approach suggested by (Lindell and Pinkas). Mapping table and graded grouping table have been used to transform data.	Applied to numerical and categorical Data sets.
Singh, Krishna, and Saxena	2009	Similarity measure metrics between two transformed points has been introduced with Jaccard similarity function and Private Equality Test (PET).	Experimental proof of proposed similarity measure is pending. Authors claim to achieve reduction in communication cost.
Zhang, Zhu, and Hua	2009	Introduced parallel algorithm with combined cryptography-based and heuristic-based approaches for privacy preservation.	Worked only for association rule mining technique by effectively concealing frequent item sets.

Table 3.4: Privacy preserving data stream mining

Chao, Sun, and Chen	2009	Data Streams Preprocessing System perturbs data streams. Online mining system uses weighted average slide window method to mine perturbed data streams.	PCDS assigns larger weights to newer data than that of older data which can better reflect current data distribution. WASW has better performance on test cases compared to VFDT.
---------------------	------	---	---

Zhang et al.	2010	Sliding window based data stream k-anonymization was proposed. It uses distribute density of data to predict upcoming data.	Proposed algorithm was tested on large anonymization value. For low density data, algorithm may not give fair performance.
Zhou et al.	2009	Author proposed k-anonymization on continuous publishing of data stream for privacy preserving using user specified maximum delay.	Clustering is used for mining on anonymized dataset. Classification model in continuous publishing data is still an open research issue.
Zhou et al.	2007	Author proposed sliding window anonymization of data stream framework (SWAF) for Privacy Protection.	SWAF is effective and efficient algorithm because it takes minor treating time for every single tuple of data steam. Privacy on data stream is still an open research issue.

3.5 Research Issues & Challenges

Present available methods are contributing privacy to the pulling out of knowledge patterns and need further exploration for possible enhancements. Common structure is still an issue that will combine more progressive measures for the estimation and the relationship of various privacy preserving data mining methodologies. In fuzzy data set, Privacy preserving data mining is still an open issues to gain privacy and at the same time to minimize the information loss. Movability data mining and privacy-aware stream data mining are among the most current and protuberant ways of privacy preserving data mining and privacy preserving data stream mining. Privacy in the context of applications, where data is releasing incrementally and in an unconditional rate is generating foremost challenges to the data mining community.

Thus, Chapter 4, 5 and 6 describe our attempt to handle following issues.

- Balancing tradeoff between privacy gain and data utility
- Various privacy preserving techniques have been developed; however, all techniques focused on only preserving the privacy and did not look into the information loss aspect.
- Accomplishing the privacy on data in data stream model is relatively hard than the traditional data mining model because of the characteristics of the data stream model.

In this chapter, we briefly presented existing privacy preserving data mining techniques which are classified into: heuristic based PPDM techniques, Reconstruction based PPDM techniques, Cryptography based PPDM techniques and PPDSM based techniques. In the next chapters, we have presented our proposed privacy preserving data stream mining algorithms which achieve privacy using perturbation and k-anonymization.

Chapter 4

Heuristic based Privacy Preserving Data Stream Mining using Hybrid Geometric Data Perturbation

In previous chapter, we have discussed relevant work for the implementation of the proposed work. In this chapter, we will provide details about our proposed algorithm that preserve privacy during the process of Data Stream Mining using Hybrid Geometric Data Perturbation.¹

4.1 Problem formulation

Here, we accept that the input consists of several nonstop streams. Without loss of generality, we may undertake that each tuple contains a single attribute. For the purpose of our examination and without loss of generality, the input containing N data streams are indicated as $D^1, D^2, D^3, D^4, \dots, D^N$. For any i^{th} data stream D^i , its value at the time t is D_t^i . The stream collection is printed as $D = [D^i \text{ for } 1 \leq i \leq N]$. Formally, the stream collection D can be considered as a $T \times N$ matrix where N is the number of streams and T is the present timestamp, which grows indefinitely. Thereafter the Hybrid Geometric Data Perturbation approach is applied on data streams. Our objective is to provide privacy before release of data streams. Per-

¹Part of this chapter has been published as Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Privacy Preserving Data Stream Mining using Hybrid Geometric Data Perturbation* in International Journal of Research in Electronics and Computer Engineering, Volume 5, PP. 107-116, 2017

turbed data streams should generate identical result as of original data stream. To preserve privacy from the available data stream, online generated noise can be addition, multiplication, and rotation. Next, mine perturbed data streams to construct a clustering model and evaluate the clustering measures. Because of data stream characteristics, Concept drift in data stream has a high level of importance especially in the case of classification. In our proposed approach concept drift is not a major concern because we are performing the clustering process using sliding window concept so the quality of information gain does not decrease.

4.2 Proposed framework

The main objective of the proposed algorithm is to provide privacy before releasing of data streams. To preserve privacy on the available data stream, online generated noise can be addition, multiplication or rotation. Then, mine perturbed data streams to construct a clustering model and evaluate the clustering measures. Our proposed work is to transform the real data set D (Stream) into modified data set D' (Stream) which is to achieve the desirable privacy on sensitive attribute data and preserve the maximum information knowledge for the intended data analysis using data mining methods. The key characteristics of our method are, that it is simple and easy to implement, less complex and requires no deep mathematical calculation. Figure 4.1 shows the framework of proposed work. Privacy-attacks to geometric data perturbations are the methods for estimating original points (or values of particular columns) from the perturbed data, with the certain level of additional knowledge about the original data. As the perturbed data goes public, the level of effectiveness is solely determined by the additional knowledge the attacker may have. Privacy preserving applications correspond to designing data management and mining algorithms in such a way that the privacy remains preserved. 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, to preserve the utility of the results. In our proposed work, sensitive attribute value is modified in such a way that attacker cannot identify the original data from perturb data and it minimizes the information loss also.

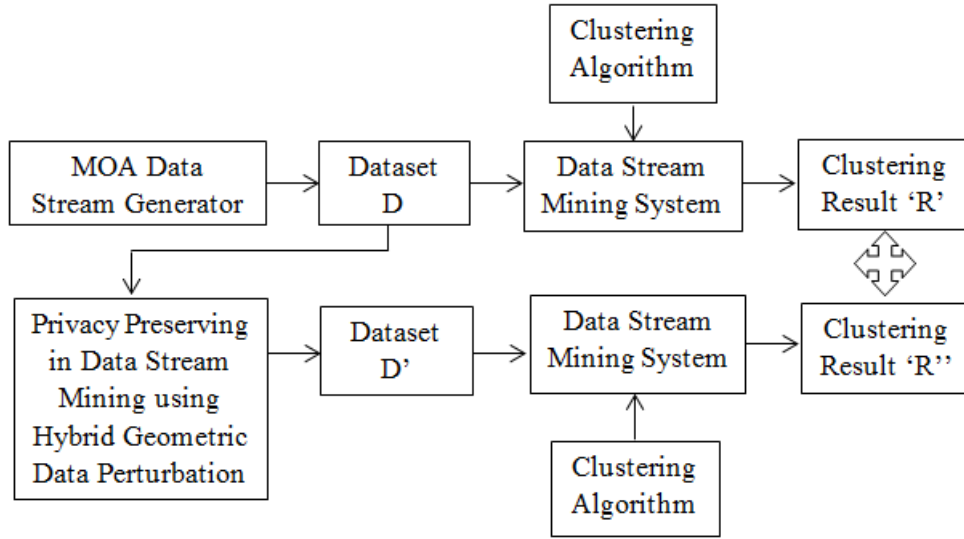


Figure 4.1: Framework of Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation

4.3 Proposed Algorithm

Our main goal is to preserve privacy in Data Stream using Hybrid Geometric Data Perturbation with 1. Minimize the information loss 2. Maximize the privacy gain and 3. Maintaining the accuracy of the clustering model. We have primarily focussed on Translation, Scaling and Rotation perturbation. We are mixing all three transformations (Translation (T), Scaling (S) and Rotation (R)) together to make a hybrid method called hybrid Geometric data perturbation. We have applied all these transformations such as TSR, TRS, STR, SRT, RTS and RST in random order. From among these orders, the one which gives us maximum privacy will be taken for granted as final procedure in the hybrid algorithm based on data set characteristics. for our algorithm, time complexity is directly proportional to the number of instances to be processed. The process is divided into two stages: a) Pre-processing on the data stream and b) Data stream analysis using clustering. The primary objective of the first stage is controlled by the data streams preprocessing system where data streams are perturbed using the proposed algorithm to protect data privacy. The primary objective of the second stage is controlled by the online data mining system which is to mine perturbed data streams to cluster the data. Proposed algorithm steps are given. Our algorithm achieves not only

Algorithm 1 Privacy Preserving in Data Stream Mining using Hybrid Geometric Data Perturbation

Input: Data Stream $D = I_1, I_2, I_3 \dots I_n$, sliding Window Size w

Intermediate Result: Perturbed Data stream $D' = I_1, I_2, I_3 \dots I_n$

Output: Clustering Result set R and R' of Data Stream D and D'

Algorithm Steps

for Each Data set D **do**

 Set SA[i] ▷ store sensitive attribute values in array

end for

for each instance I in Data Set D **do**

for $i=0$ to w **do**

 SA[i]=TDP() ▷ TDP-Translation Data Perturbation

 //TDPSteps

for each confidential attribute A_j in D , where $1 \leq j \leq d$ **do**

 Select the noise term e_j in N for the confidential attribute A_j

$$A'_j = A_j + e_j$$

end for

 SA[i]=SDP() ▷ SDP-Scaling Data Perturbation

 //SDPSteps

for each confidential attribute A_j in D , where $1 \leq j \leq d$ **do**

 Select the noise term e_j in N for the confidential attribute A_j

$$A'_j = A_j * e_j$$

end for

 SA[i]=RDP() ▷ Rotation Data Perturbation

 //RDPSteps

$$k \leftarrow \lfloor n/2 \rfloor$$

$p_k \leftarrow k$ Pairs(A_i, A_j) in D such that $1 \leq i, j \leq n$ and $i \neq j$

for each selected pair P_k in Pairs(D) **do**

$$V(A'_i, A'_j) \leftarrow R_\theta * V(A_i, A_j) \quad \triangleright V \text{ is computed as a function of } \theta$$

$$\theta_k \leftarrow \text{SecurityRangevalueof } \theta_k(30, 45, 60, 90)$$

$$V(A'_i, A'_j) \leftarrow R_{\theta_k} * V(A_i, A_j) \quad \triangleright \text{Output the distorted attributes of } D$$

```

for each algorithm of TDP,SDP and RDP) do
     $Noise(N) = Average(Attributes\ except\ sensitive\ attributes)$ 
end for
end for
end for
Store Perturbed Dataset in  $D'$ 
Apply  $K - means$  clustering algorithm on  $D'$ 
end for

```

privacy with minimum information loss but it also finds out the transformation orders that giving maximum privacy. Proposed approach does not consider overall statistical characteristic of data sets, Instead it only focuses on attribute that is to be protected. Sliding window based hybrid geometric based data perturbation method is very simple and speeds up the data perturbation task. Unlike many other data perturbation techniques, which concentrated on balancing the tradeoff between the level of data utility and data privacy, the proposed algorithm concentrates on how to maximize the desired data privacy. All three transformations are describe below:

-Translation Data Perturbation

A constant or noise is added into sensitive attribute values. The constant can be a positive or negative number. Although its degree of privacy protection is '0' in accordance with the formula for calculating the degree of privacy protection, we cannot see the raw data from transformed data directly.

-Scaling Data Perturbation

It is also known as ratio transformation. In the ratio transformation, x-axis, y-axis are fixed straight lines. Scaling transformation is achieved by multiplying a constant or noise to all the values of an attribute. The constant can be positive or negative number.

-Rotation Data Perturbation

For a pair of attributes arbitrarily chosen, consider them as points of two dimension space, and rotate them according to a given angle θ with the origin as the center. If θ is positive, we rotate them along the anti-clockwise. otherwise we rotate

them along the clockwise. Rotation can be achieved by multiplying the matrix:

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

We extended existing MOA framework in which, each tuple of data stream is independently treated. We considered single attribute as sensitive attribute (dependent attribute) and rest are as non-confidential attributes (independent attributes) ignoring class attribute. We focused on finding noise which will help us protect our sensitive attribute value while maintaining the accuracy of clustering result and evaluation measures as we have in true dataset. The value of sensitive attribute of each tuple is modified independently from other tuples. In order to find noise by what the sensitive attribute is protected, the nonconfidential attributes were taken into account. Entire flow of the proposed work is shown in figure 4.2.

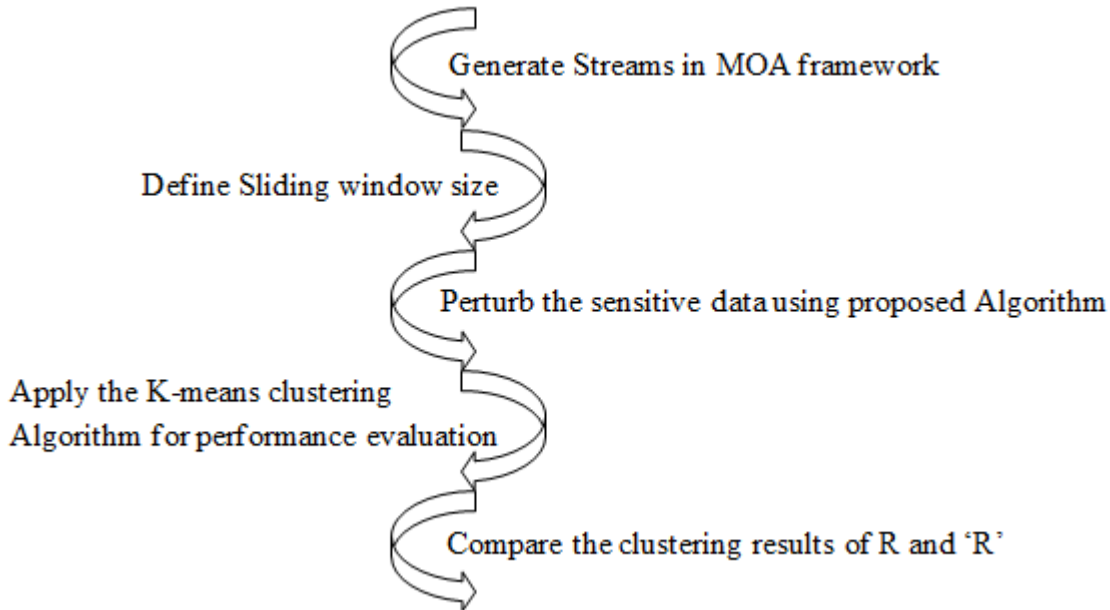


Figure 4.2: Flow of the proposed Framework

4.4 Performance Evaluation

As per our previous discussion, proposed approach is divided into two stages; 1) Data stream preprocessing 2) Data stream cluster mining. In the first stage of data stream preprocessing, upon receiving data stream from data stream generator or

real dataset, the Data Stream Preprocessing System (DSPS) uses perturbation algorithm to perturb confidential data. We can flexibly adjust the data attribute to be perturbed according to the security need. Therefore, threats and risks from releasing data can be effectively reduced. In the second stage of data stream mining, the online data mining system uses the sliding window mechanism to cluster perturbed data streams. Experimental results show that the proposed approach not only preserves data privacy but also mines data stream accurately. Series of experiments were performed over defined sliding window size (w) in order to evaluate the clustering accuracy. Our evaluation approach focused on the overall quality of generated clusters after dataset perturbation.

4.4.1 Experimental setup

To evaluate the effectiveness of proposed privacy preserving method, experiments have been carried out on Intel Core I3 Processor with 3 GB primary memory on Windows system. Simulation has been done in data stream clustering environment. We quantified proposed approach using resultant accuracy of true dataset clustering and perturbed dataset clustering. Proposed approach has been implemented in Java and integrated with MOA framework. Experiments are performed based on sliding window size (w) concept in order to estimate the clustering accuracy. The presented work is dedicated to the entire quality of produced clusters after dataset perturbation. Data set D is given as an input to proposed data stream perturbation algorithm. Algorithm perturbs only sensitive attribute values and resultant dataset with modified values is called perturbed dataset D' . D and D' are provided to standard clustering stream learning algorithms to obtain results R and R' respectively. Proposed work focuses on obtaining close approximation between clustering results R and R' to balance in between privacy improvement and information damage. The primary objective of the second stage, which is handled by the online data mining system, is to mine perturbed data streams to cluster the data. K-Mean clustering algorithm over predefined sliding window size on perturbed data stream has been used in order to measure the exactness and usefulness of clustering outcomes over five different ordinary datasets (Coverttype, Electricity, Agrawal, Bank Marketing, Airlines). Outcomes show that properly best level

of privacy has been accomplished with reasonable accuracy in almost all test cases. Accuracy between original dataset and perturbed dataset has been quantified by percentage of instances assigned to different clusters with the help of cluster membership matrix (CMM). Proposed approach shows reasonably good results against evaluation measures Precision, Recall, Misclassification and CMM (Cluster Membership Matrix). With the help of CMM, we matched how closely each cluster in the perturbed dataset equals its equivalent cluster in the original Dataset. We intend to use such a matrix as the Clustering Membership Matrix (CMM) where the row show the clusters in the original dataset, the columns represent the clusters in the perturbed dataset. We constructed each dataset to define five clusters using k-Mean clustering procedure. Each matrix demonstrates five clusters situation for real dataset and perturb dataset. Real dataset clustering outcome gives information about number of occurrences actually classified in each cluster whereas perturbed dataset clustering shows outcome of correct assignments after data perturbation and percentage of accuracy accomplished. Table 4.2 to 4.4 show the Membership Matrix with best Information Gain (Accuracy) after performing Geometric Perturbation on various data streams with different window size. K-Mean Clustering algorithm using WEKA data mining tool in MOA framework has been simulated to evaluate the accuracy of proposed PPDSM approach. MOA is the tool for implementing methods and running experiments for online learning from evolving data streams (Bifet et al.). MOA supports evaluation of data stream learning algorithms on large streams for both Clustering and Classification. In addition to this, it also supports interface with WEKA machine learning algorithms. Following steps of MOA framework describe how data stream mining with proposed perturbation technique works.

- Structure each dataset as streaming data in MOA framework.
- Define sliding window (W) over the data stream.
- Apply our proposed method to protect the sensitive attribute value to achieve privacy on dataset.
- Apply the K-mean method to find the clusters of perturbed dataset and original dataset. K-mean is scalable and known method which is used on static

dataset and streaming data.

- Match the clusters of perturbed dataset with clusters of original dataset. F-measure is useful to measure the quality of clusters.

4.4.2 Experimental Results

Result evaluation mainly focusses on the overall quality of generated clusters after data stream perturbation. We compare how closely each cluster in the perturbed data set matches its corresponding cluster in the original data set. To do so, we first identify the matching of clusters by computing the matrix of frequencies.

Table 4.1: Clustering Membership Matrix(CMM)

	C1'	C2'		Cn'
C1	$Freq_{1,1}$	$Freq_{1,2}$	$Freq_{1,n}$
C2	$Freq_{2,1}$	$Freq_{2,2}$	$Freq_{2,n}$
:	:	:	:
Cn	$Freq_{n,1}$	$Freq_{n,2}$	$Freq_{n,n}$

We refer to such matrix as Clustering Membership Matrix (CMM) shown in table 4.1, where the rows represent the clusters in the original dataset, the columns represent the clusters in the perturbed dataset, and $Freq_{i,j}$ is the number of points in cluster C_i that falls in cluster C'_j in the perturbed dataset. After computing the frequencies $Freq_{i,j}$, we scanned the CMM to calculate percentage of accuracy of perturb data set for each cluster C'_j with respect to C_i in the original dataset. After generating this CMM table, we have calculated information loss and information gain using following equation.

$$InformationGain = \frac{(\sum_{i=1}^n (C'x100))}{\sum_{i=1}^n (C)}$$

$$InformationLoss = InformationGain(C) - InformationGain(C')$$

Table 4.2 to 4.4 show the result of Cluster Membership Matrix with maximum Information Gain after performing Geometric Perturbation on various data streams with different window size. Table 4.2 shows the information gain result on cover-type data set where we set the window size (w) to 1000, Rotation angle 30 degree

and order of transformation is Translation, Scaling and Rotation (TSR). We have applied the k-means algorithm where the value of $k=5$ on original data set and perturbed data set. According to this table, information gain is 87.02% and information loss is 12.98%. Table 4.3 and 4.4 also show the information gain result with different window size (w), rotation angle and order of transformation. According to table 4.3 and 4.4, information gain is 86.68% and 83.04%. The experiments are processed on five different data sets obtainable from the UCI Machine Learning Repository, MOA dataset dictionary and Agrawal Weka dataset. We restricted our experiment to numeric attributes; even we can extend the implementation to categorical attributes. We performed our experiments on Covertypes, Electricity, Agrawal, Bank Marketing, and Airlines dataset. Details about these data sets are given in Appendix-A.

Table 4.2: Information Gain for Covertypes Dataset ($W=1000$, Angle = 30° , Sequence = TDP_SDP_RDP)

Clusters	Original Dataset	Perturbed Dataset	Information Gain (%)
	Number of Instances per Cluster	Number of Instances per Correctly Clustered	
C1	14240	12548	88.12
C2	13440	11465	85.30
C3	12691	11610	91.48
C4	11788	09761	82.80
C5	12841	11229	87.44
Total	65000	56613	87.02

Table 4.3: Information Gain for Coverttype Dataset (w=3000, Angle = 30°, Sequence = TDP_SDP_RDP)

Clusters	Original Dataset	Perturbed Dataset	Information Gain (%)
	Number of Instances per Cluster	Number of Instances per Correctly Clustered	
C1	14984	12683	84.64
C2	13803	11776	85.31
C3	10710	09621	89.83
C4	11776	09659	82.02
C5	11727	10870	92.69
Total	63000	54609	86.68

Table 4.4: Information Gain for Coverttype Dataset (w=5000, Angle = 60°, Sequence = RDP_SDP_TDP)

Clusters	Original Dataset	Perturbed Dataset	Information Gain (%)
	Number of Instances per Cluster	Number of Instances per Correctly Clustered	
C1	16196	13422	82.87
C2	12441	11938	95.96
C3	12376	11084	89.56
C4	11720	09019	76.95
C5	12267	08515	69.41
Total	65000	53978	83.04

Table 4.5 show the Information gain after applying our proposed approach on Coverttype perturbed dataset with different angle and different order of transformations. Figure 4.3 to 4.17 show the information gain result using various transformations in MOA framework. We have applied our proposed approach on five different datasets with different Sliding window size and rotation angle. We can conclude that TSR, TRS and STR based Geometric Process is more suitable to achieve maximum privacy with minimum information loss. Among these

Table 4.5: Information Gain (Covertypes Dataset)

Dataset	Sliding Window Size (w)	Angle	TSR (%)	TRS (%)	STR (%)	SRT (%)	RTS (%)	RST (%)
Covertypes	1000	30°	87.09	87.09	87.26	81.84	82.1	81.84
		45°	83.8	83.8	83.53	82.82	82.69	82.82
		60°	82.75	82.75	82.82	82.9	82.9	82.91
		90°	81.92	81.92	81.35	84.86	83.84	84.86
	3000	30°	86.68	86.68	89.35	83.93	82.81	83.93
		45°	82.73	82.73	82.79	83.65	82.9	83.65
		60°	82.89	82.89	82.93	84.07	83.56	84.07
		90°	83.43	83.43	81.79	84.42	82.85	84.42
	5000	30°	82.55	82.55	82.64	82.65	81.78	82.65
		45°	81.12	81.12	81.29	82.04	82.15	82.04
		60°	82.12	82.12	81.59	83.04	82.87	83.04
		90°	79.16	79.16	80.25	78.331	78.39	78.31

three, user or organization can select any one method for preserving privacy with minimum information loss.

In proposed work, analysis of Accuracy is to evaluate the clustering measures with the help of MOA framework. We concentrated on two essential measures F1_P (determine the precision of system by considering the precision of individual cluster) and F1_R (determine the recall of system, which take into account the recall of each cluster). Results are presented in terms of graphs where each graph contains the measure we obtained when original data is processed without applying privacy preserving method and when data undergoes through our proposed privacy preserving method. Occurrences are treated in defined sliding window size. Number of result graph is generated. Figure 4.18 to 4.27 show the Precision (Calculated using MOA equation) and Recall (Calculated using MOA equation) graph which measures the accuracy.

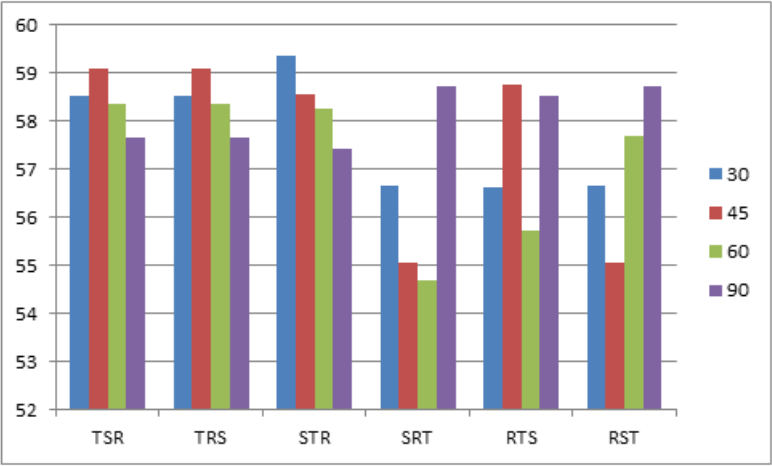


Figure 4.3: Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 1000)

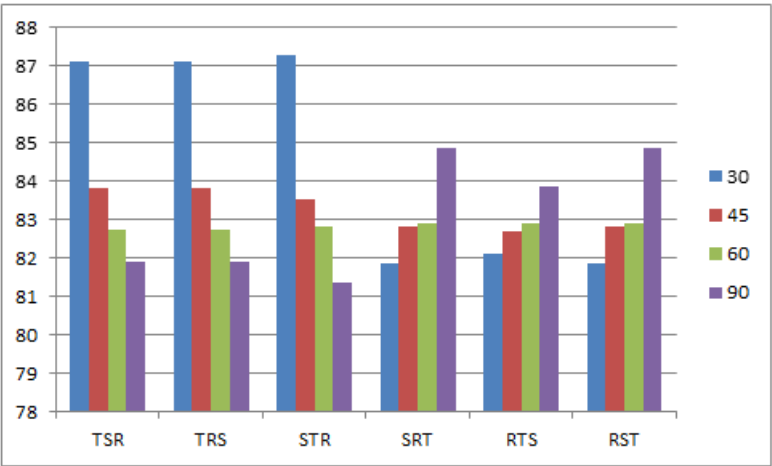


Figure 4.4: Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 1000)

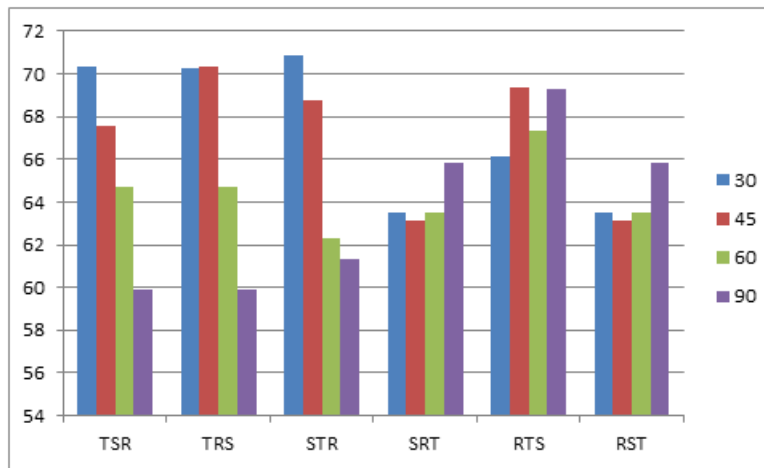


Figure 4.5: Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, $w = 1000$)

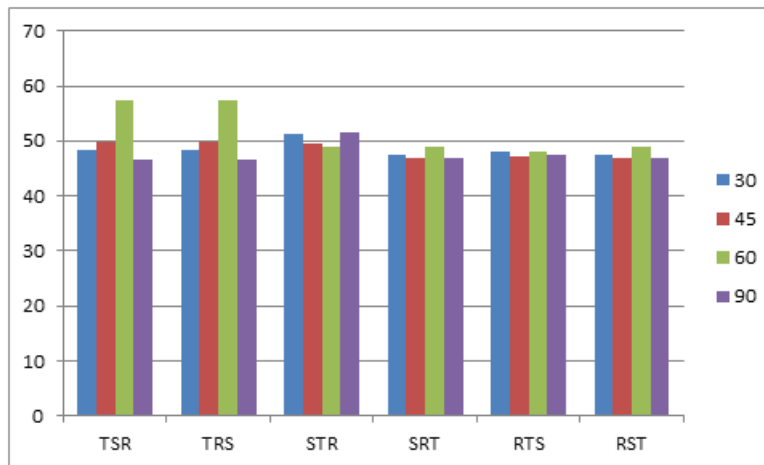


Figure 4.6: Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, $w = 1000$)

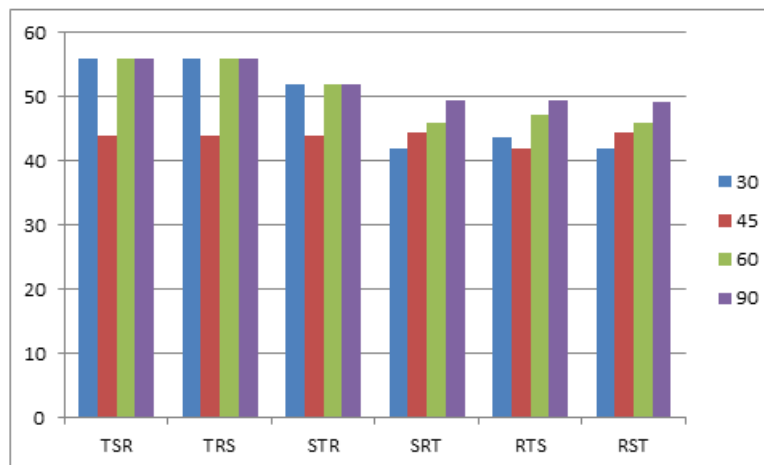


Figure 4.7: Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, $w = 1000$)

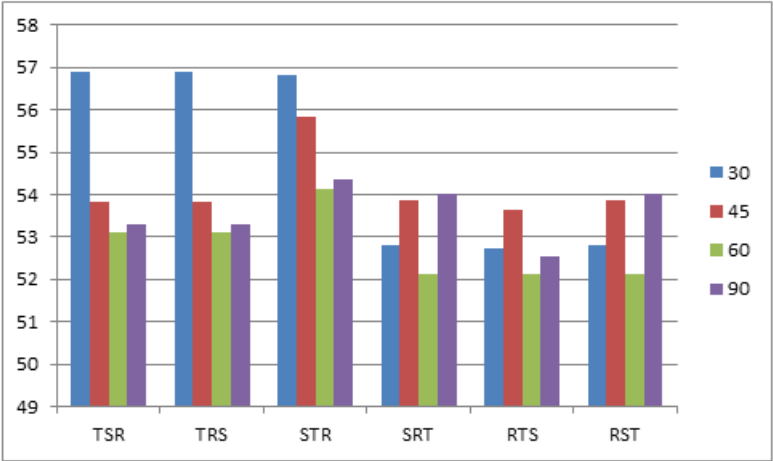


Figure 4.8: Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 3000)

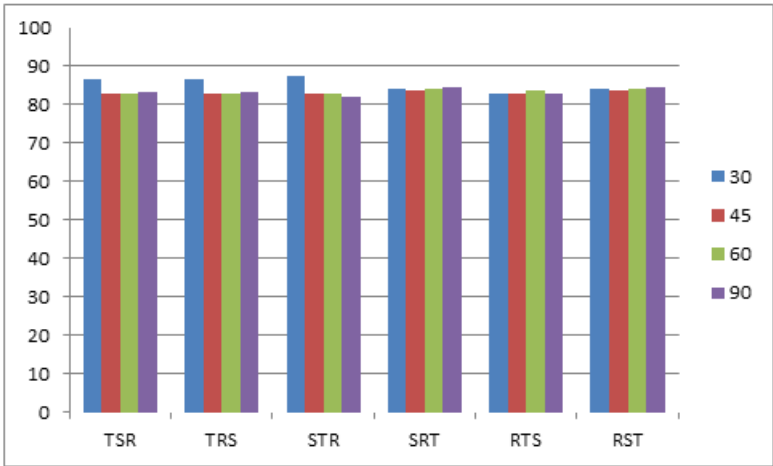


Figure 4.9: Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 3000)

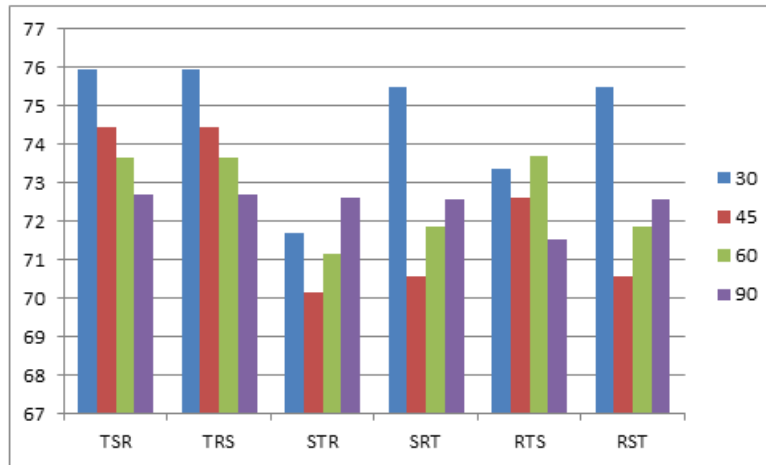


Figure 4.10: Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, $w = 3000$)

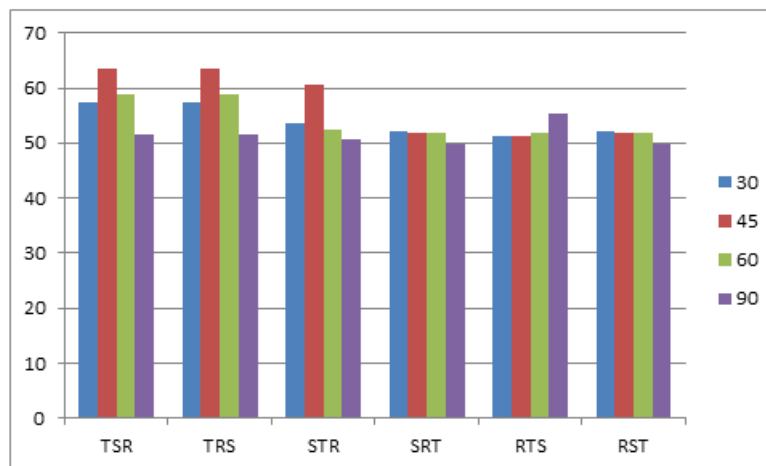


Figure 4.11: Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, $w = 3000$)

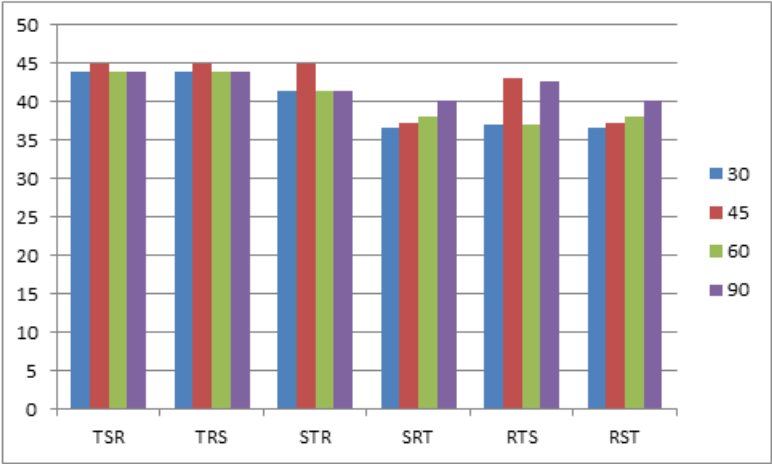


Figure 4.12: Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, w = 3000)

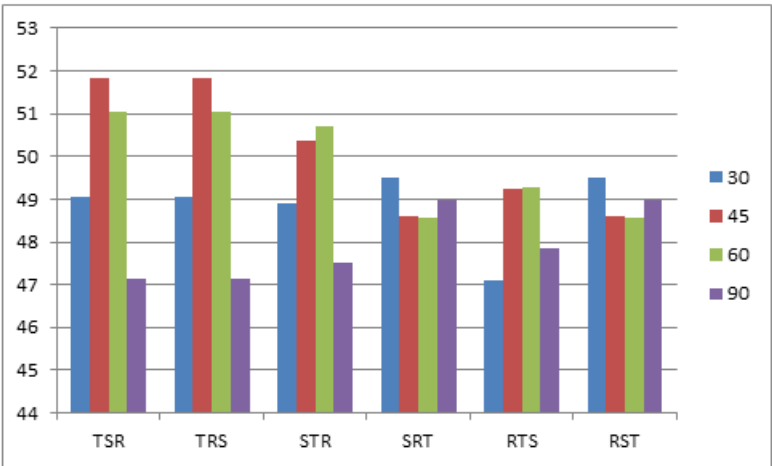


Figure 4.13: Accuracy measurement (Agrawal dataset, Angle = 30,45,60,90, w = 5000)

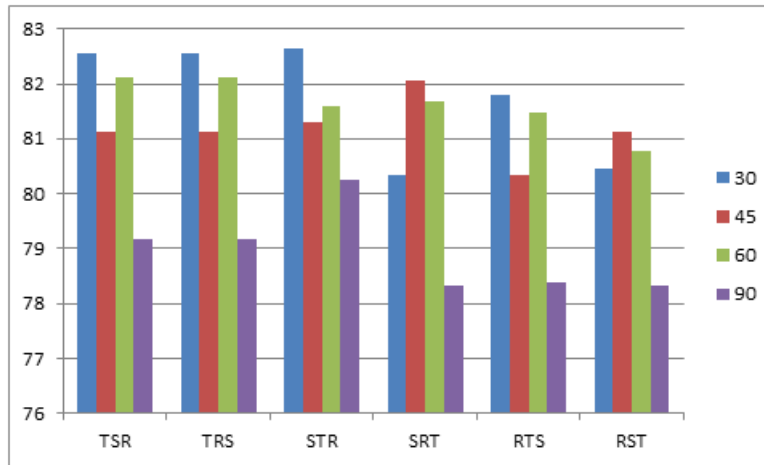


Figure 4.14: Accuracy measurement (Covertypes dataset, Angle = 30,45,60,90, w = 5000)

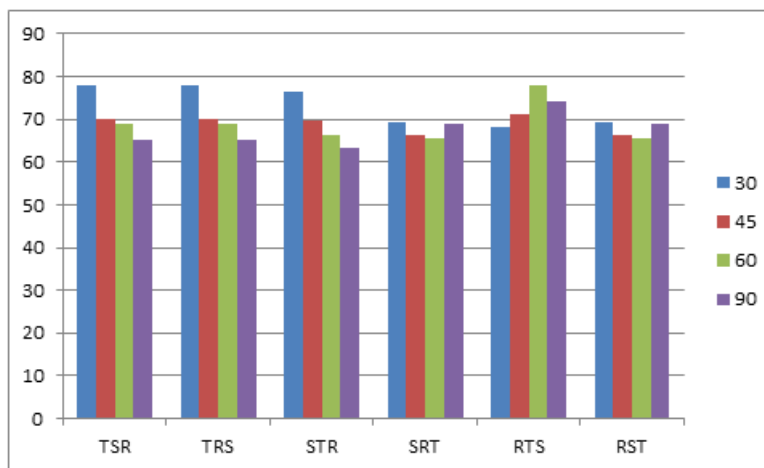


Figure 4.15: Accuracy measurement (Electrical dataset, Angle = 30,45,60,90, w = 5000)

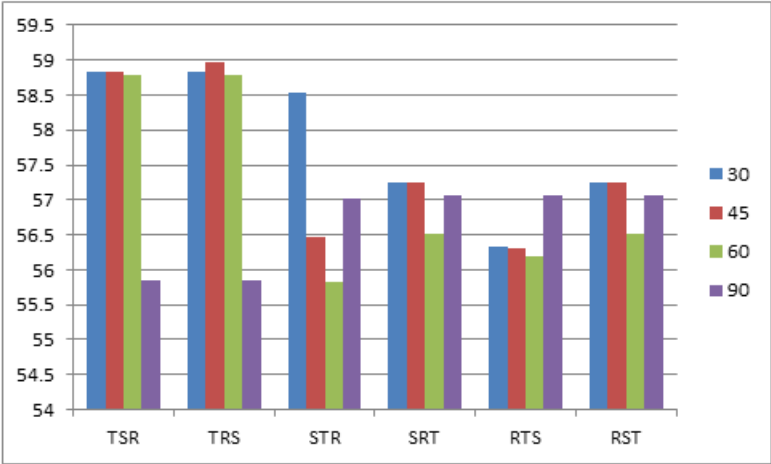


Figure 4.16: Accuracy measurement (Bank marketing dataset, Angle = 30,45,60,90, w = 5000)

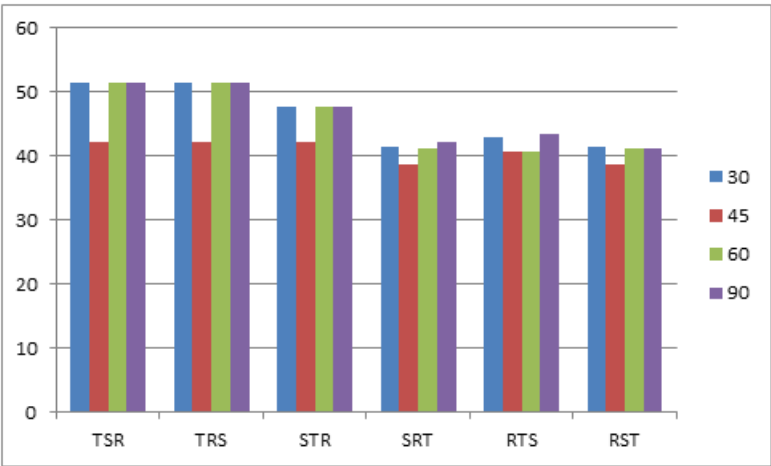


Figure 4.17: Accuracy measurement (Airlines dataset, Angle = 30,45,60,90, w = 5000)

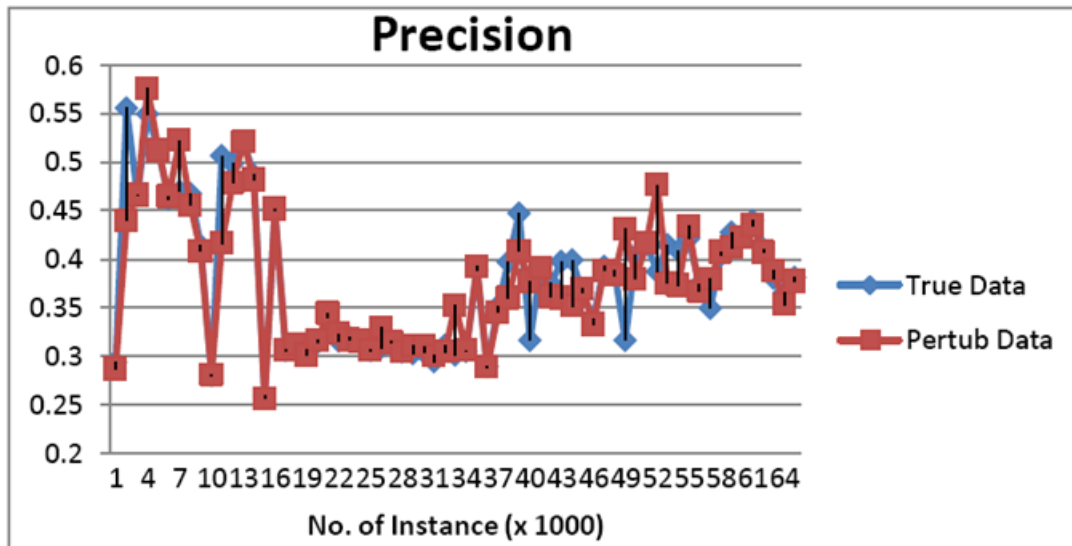


Figure 4.18: Accuracy on Covertypes Dataset ($w = 1000$, Angle = 30, Sequence = SDP_TDP_RDP)

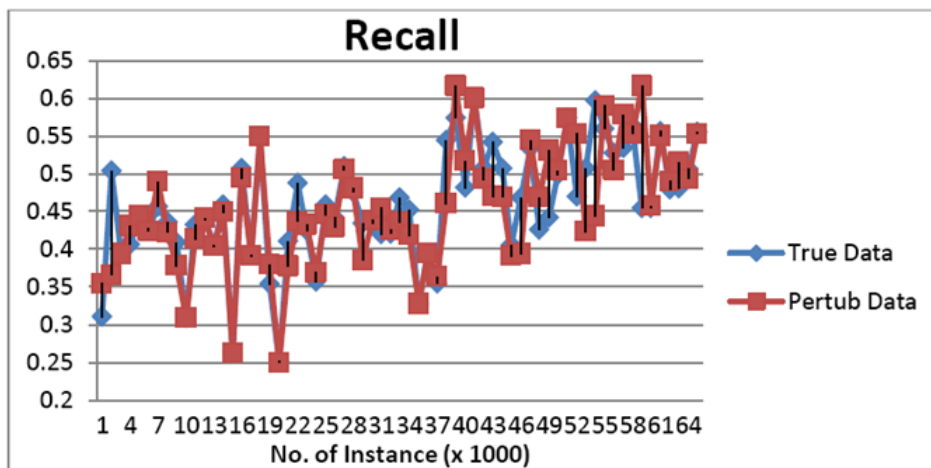


Figure 4.19: Accuracy on Covertypes Dataset ($w = 1000$, Angle = 30, Sequence = SDP_TDP_RDP)

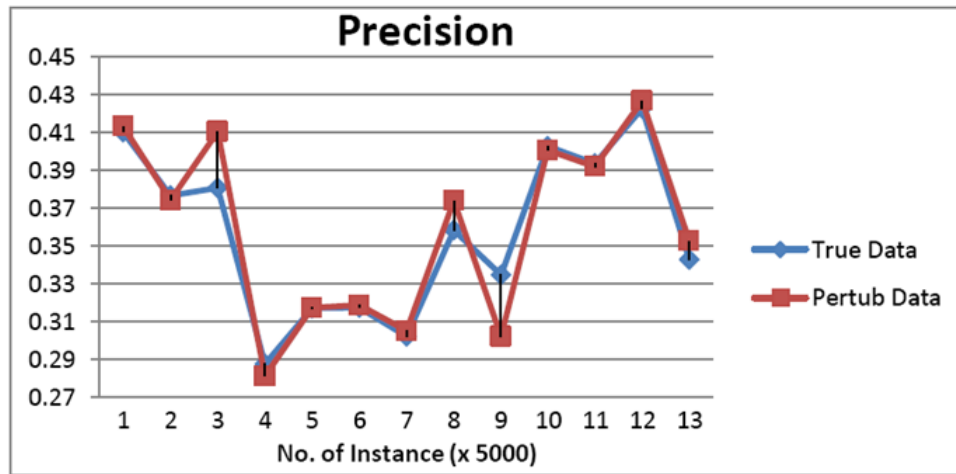


Figure 4.20: Accuracy on Coverttype Dataset ($w = 5000$, Angle = 60, Sequence = SDP_TDP_RDP)

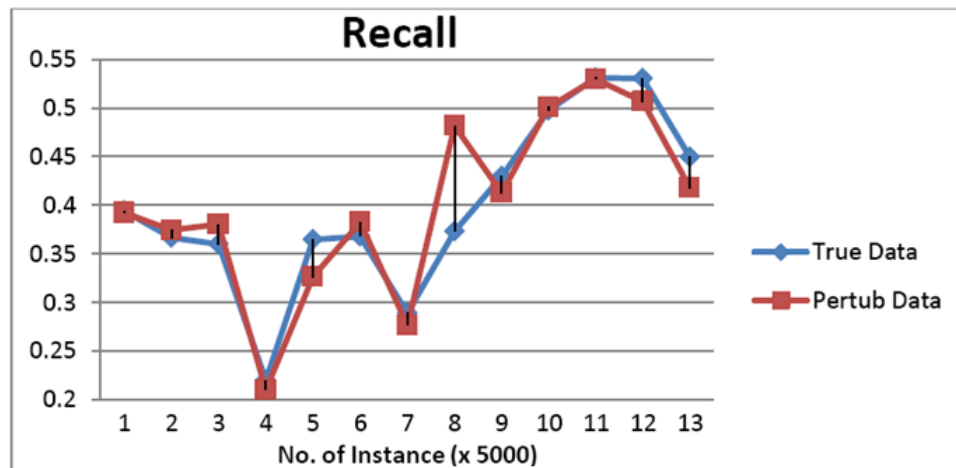


Figure 4.21: Accuracy on Coverttype Dataset ($w = 5000$, Angle = 60, Sequence = SDP_TDP_RDP)

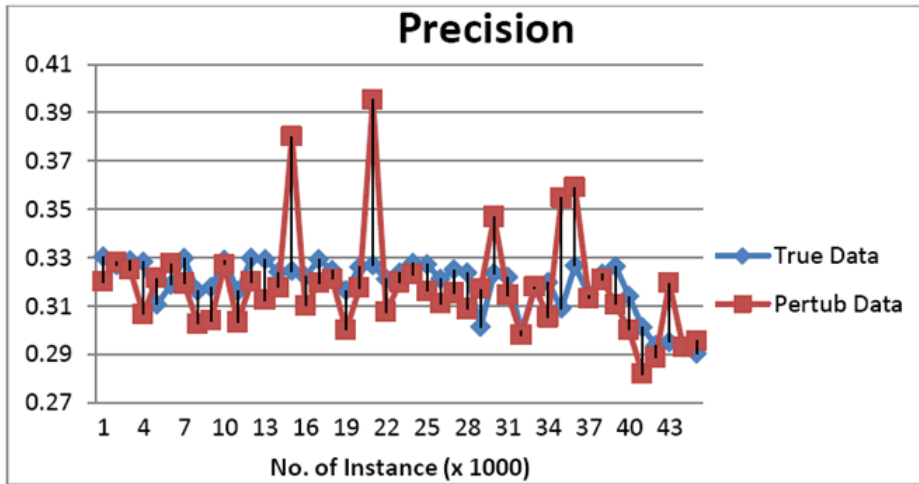


Figure 4.22: Accuracy on Bank Dataset ($w = 1000$, Angle = 60, Sequence = SDP_TDP_RDP)

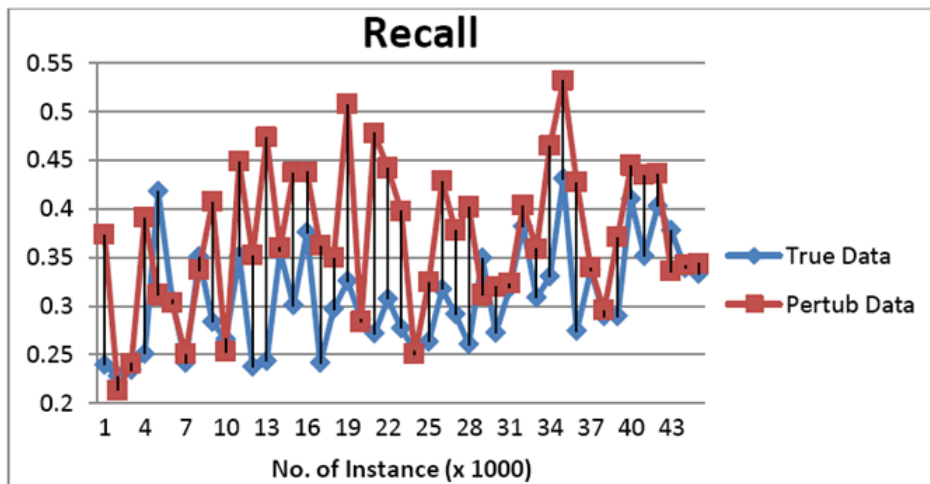


Figure 4.23: Accuracy on Bank Dataset ($w = 1000$, Angle = 60, Sequence = SDP_TDP_RDP)

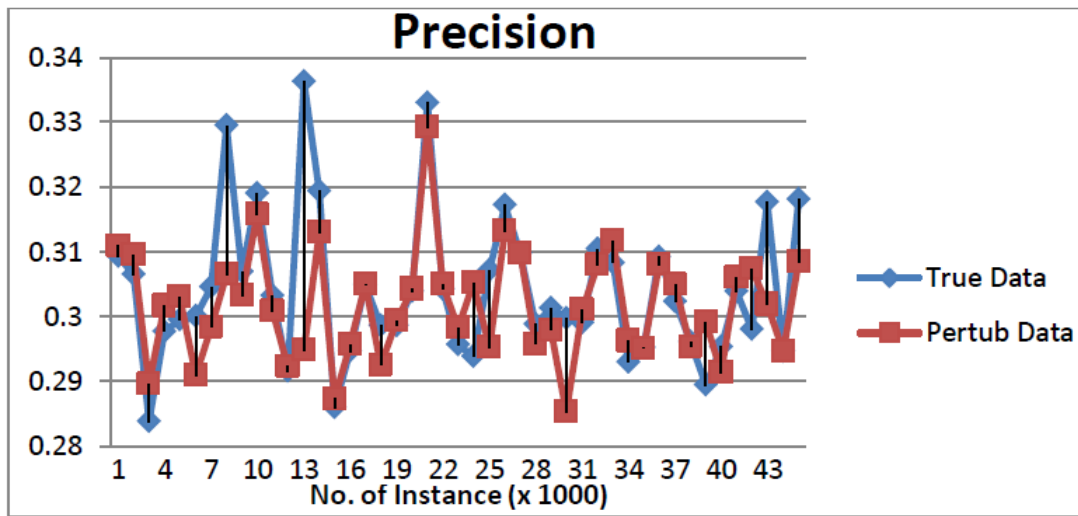


Figure 4.24: Accuracy on Electricity Dataset ($w = 1000$, Angle = 30, Sequence = SDP_TDP_RDP)

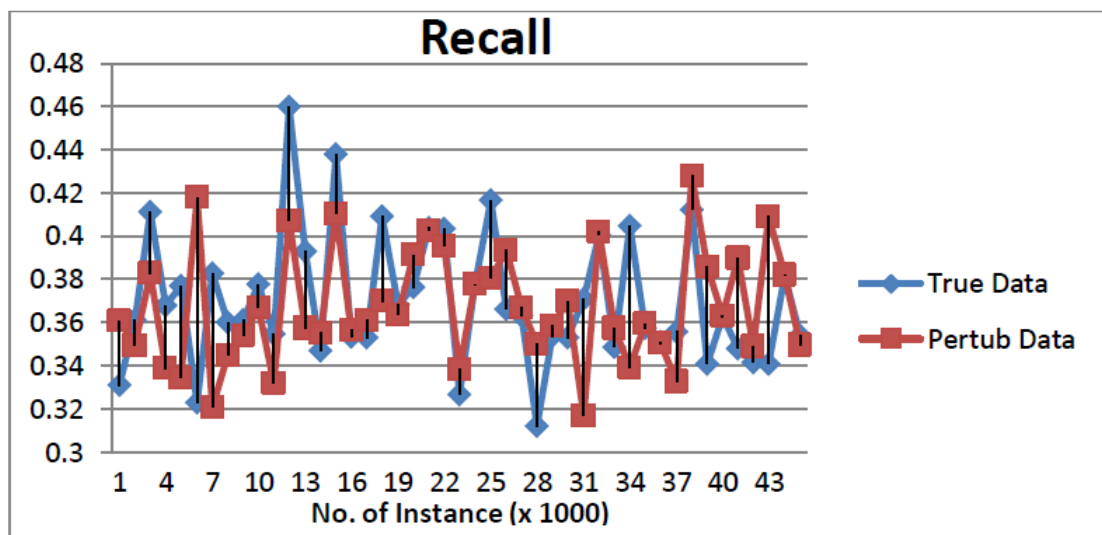


Figure 4.25: Accuracy on Electricity Dataset ($w = 1000$, Angle = 30, Sequence = SDP_TDP_RDP)

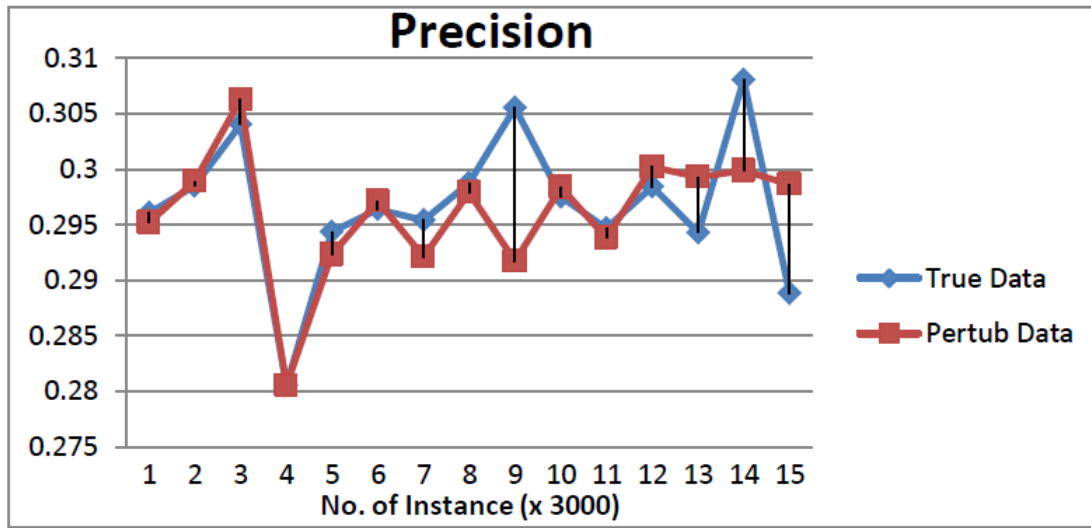


Figure 4.26: Accuracy on Electricity Dataset ($w = 3000$, Angle = 60, Sequence = SDP_TDP_RDP)

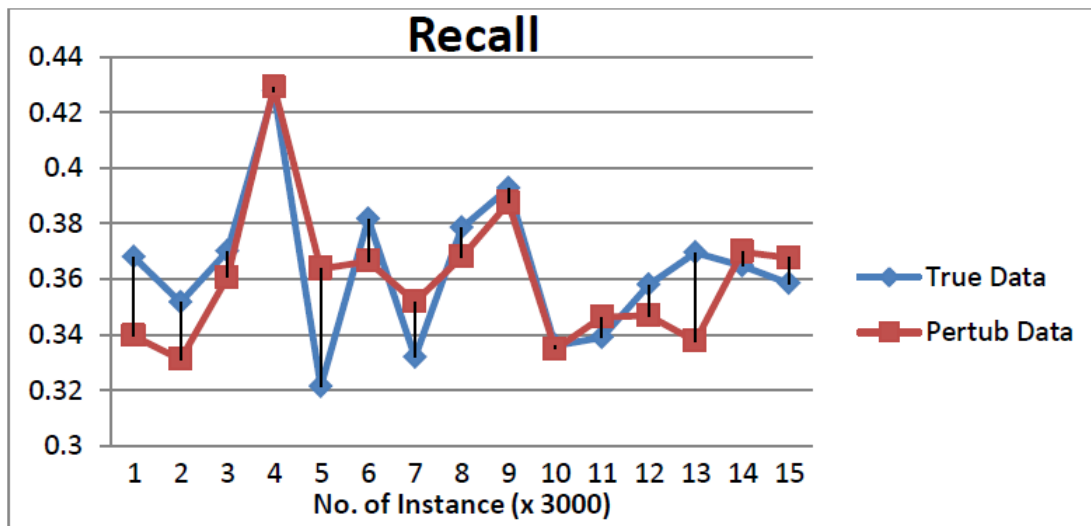


Figure 4.27: Accuracy on Electricity Dataset ($w = 3000$, Angle = 60, Sequence = SDP_TDP_RDP)

4.4.3 Comparison of Proposed Approach with GDP Approach

A comparison of results of proposed approach with existing Geometric Data Perturbation approach (Chen and Liu, "Privacy-preserving multiparty collaborative mining with geometric data perturbation") with same number of records of given data set is depicted in figure 4.28. Graph show the average result of accuracy (infor-

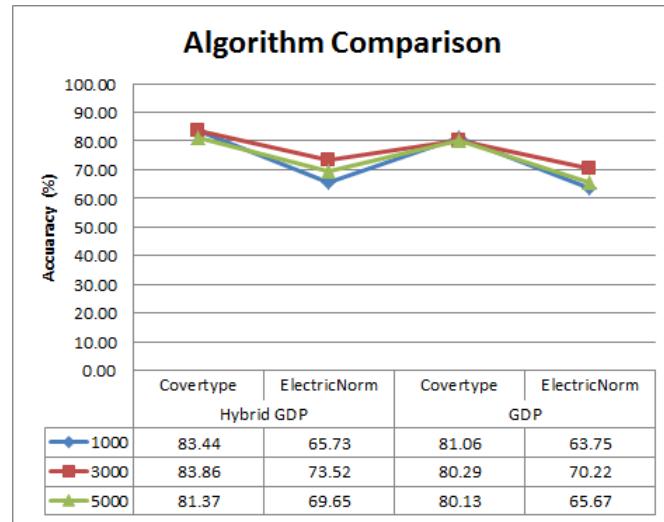


Figure 4.28: Comparison of proposed approach with GDP

mation gain) measurement of covertypes and ElectricNorm data set using different sliding window size. Results shows that proposed approach performs better in comparison with existing GDP based approach.

4.5 Summary

In this chapter, we have proposed heuristic based Privacy Preserving Data Stream Mining using Hybrid Geometric Data Perturbation approach. Perturbation approaches assure that exposure will not occur. Revealing the data from perturbation depends on algorithm strength. We have found that if sensitive attributes independently perturb then it is possible that complete or partial data is disclosed. Most of the reconstruction perturbation based methods are disclosed, if attacker have the knowledge about methods. Therefore, it is compulsory to measure the level of security provided by a specific perturbation method when quantifying privacy by such a method. Undoubtedly, the above measure to quantify privacy is based on how closely the original values of altered values are estimated. Our proposed

approach maintains the variance level in such a manner that it may not be possible that adversary can disclose the original data back from perturbed data. In the next chapter, we have presented another privacy preserving data stream mining approach which achieves the privacy using sensitive drift (SD) on sensitive attribute value.

Chapter 5

SD – Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering

In the previous chapter, we proposed geometric based data perturbation algorithm for preserving privacy on sensitive attribute values. In this chapter, we have proposed algorithm that preserve privacy during the process of data stream mining and balance trade off with data utility. Statistical characteristics of data set have been considered before applying perturbation and values have been changed keeping such characteristics in place. Proposed SD -Perturbation: A Sensitive Drift based perturbation algorithm for privacy preserving in Data Stream Clustering method has been used sensitive drift value for data perturbation and then, evaluated with Precision, Recall, Bias in Mean (BIM), Bias in Standard Deviation (BISD) and Correlation parameters over clustering algorithm. ¹

¹Part of this chapter is communicated as Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *SD Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering* in International Journal of Information and Communication Technology, inderscience, 2018

5.1 Problem formulation

There are number of government sectors and public sectors, where digital data can rise without bound at a high rate (e.g. millions) per day and needs to be processed in real time with minimal delay. Let $DStream$ be the data stream, t be the time stamp and w be the defined window size. Let m be the number of attributes into each data streams. Each data stream at different time stamps is represented as;

Data stream $DStream_1$: $DStream_1^{t1}, DStream_1^{t2}, DStream_1^{t3}, \dots, DStream_1^w$

Data stream $DStream_2$: $DStream_2^{t1}, DStream_2^{t2}, DStream_2^{t3}, \dots, DStream_2^w$

Data stream $DStream_3$: $DStream_3^{t1}, DStream_3^{t2}, DStream_3^{t3}, \dots, DStream_3^w$

.....
.....

Data stream $DStream_n$: $DStream_n^{t1}, DStream_n^{t2}, DStream_n^{t3}, \dots, DStream_n^w$

Each data streams $DStream_{w \times m}$ can be mined using existing data stream clustering algorithms.

5.1.1 Privacy in Data Stream Mining

Provide the privacy on data stream before it is publicly available for mining purpose. Perturbed data set ($\sim DStream_{w \times m}$) should generate identical result of mining as of original data stream. To preserve the privacy on sensitive data stream values, online generated noise (N) can be added / multiplied (\oplus) in original data stream ($DStream$).

$$DStream_{w \times m} \oplus N_w = \sim DStream_{w \times m}$$

Figure 5.1 shows that, noise addition / multiplication is based on i. i. d. (Identical independently distribution) in data stream.

5.1.2 Privacy in Data Mining

Privacy is applied on sensitive data values before publishing large static data set for mining purpose. It is necessary that, perturbed data set should produce identical result as of original data set during the process of data mining. To preserve privacy on sensitive data values from available static data set, noise is produced and added / multiplied with original Data set. Generated noise is based on a prob-

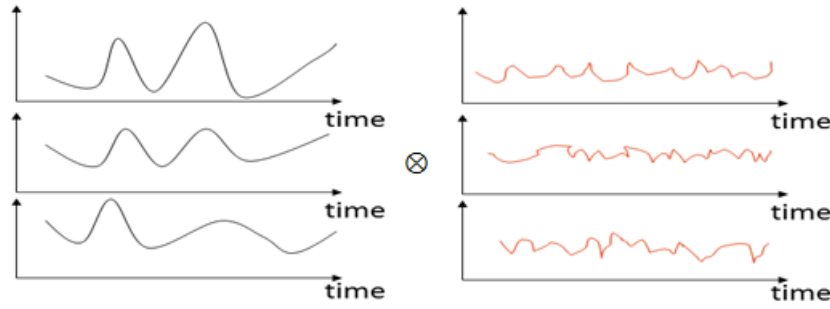


Figure 5.1: Identical independently distributed noise addition into original data streams

ability density function that understands the statistical characteristics of original data set.

$$Dataset \oplus N = \sim Dataset$$

Figure 5.2 shown Principal Component Analysis (PCA) based data reconstruction. Where, D is an original data set which will be changed to D^* after applying data perturbation algorithms. D^\sim is the possible reconstruction of data from D^* . Privacy gain is to protect original data set D to be reconstructed from perturbed data set

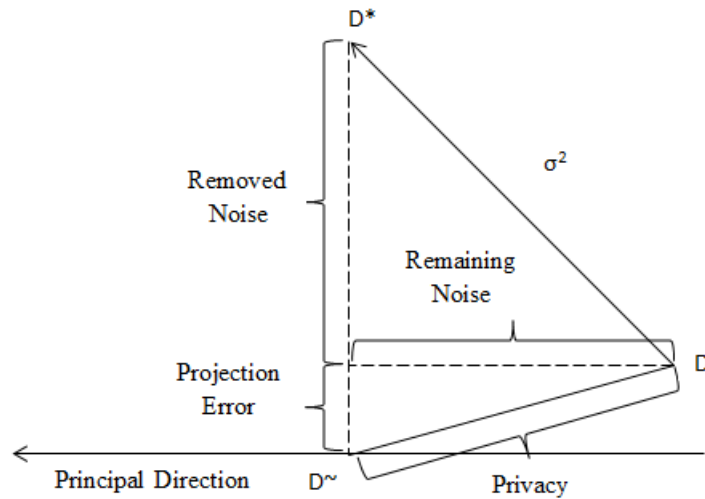


Figure 5.2: Principal Component Analysis (PCA) based data reconstruction

5.2 Proposed framework

(Chhinkaniwala and Garg) proposed the method to protect the sensitive attribute value via tuple value (tuple value is the average of Z-score normalized values of

attributes of given instance except the class attribute). The tuple values are then multiplied with the values of sensitive attribute of respective instances. Existing approach for tuple value computation using normalized values of attributes (except sensitive attribute) has been compared with other tuples having the same normalized values. However, it is possible that such tuple could have the same tuple value as the other. In such a scenario, perturbation of a sensitive attribute value may compromise the privacy of certain set of tuples. To overcome the above mentioned issue, a sensitive drift has been introduced. Sensitive drift approach does not claim to improve privacy gain in all cases, but it provides more robustness to the existing approach. The overall objective of our work is to enhance privacy in data stream mining with minimum information loss. The proposed work satisfies the properties like, maintain the statistical properties on data set after applying privacy on original data set, sensitive attribute values hide the relationship between the other attributes of the data set and sensitive drift to achieve the privacy on sensitive attributes. Figure 5.3 shows the entire process of proposed work.

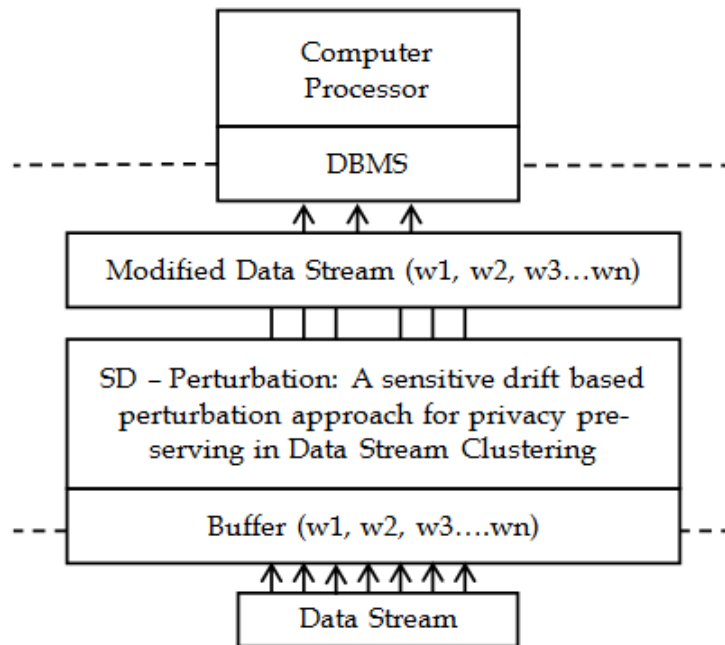


Figure 5.3: Entire process of SD Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering

Proposed work is divided into two stages: data stream preprocessing and clustering in data stream mining. The primary objective of the first step is to per-

turb data streams to preserve data privacy. Using the sliding window concept, data streams are converted into normalized forms (z-score or min-max normalization). Calculate the tuple value using tuple value function, which is the average of the normalized values. Next, based on user-defined window size (number of records) and sensitive drift value (in percentage), find out the upper bound and lower bound of tuple values for current instance of tuple value as per algorithm. Select the tuple values which are between the range of upper bound and lower bound. Select those sensitive attribute values which are mapped with tuple values that were selected previously. Find out the average of these sensitive attribute values and replace current instance of sensitive attribute value with the average of these sensitive attribute values. Repeat these steps for whole data stream, and finally, we will get perturbed data streams. The primary objective of the second stage is to perform the cluster mining on perturb data set to measure the performance using cluster membership matrix (CMM) and precision/recall method.

In first stage of proposed algorithm, Dataset D is given as an input to proposed data perturbation algorithm. Algorithm perturbs only sensitive attribute values and resultant dataset with modified values which is called perturbed dataset D' . D and D' are provided to standard clustering stream learning algorithms to obtain results R and R' respectively. The proposed work focusses on obtaining close approximation between clustering results R and R' to balance tradeoff between privacy gain and information loss. The primary objective of the second stage, which is handled by the online data mining system, is to mine perturbed data streams to cluster the data. Figure 5.4 shows the proposed framework which is the extended framework of MOA.

Proposed work focuses on obtaining close approximation between clustering results R and R' to balance trade-off between privacy gain and information loss. The primary objective of the second stage, which is handled by the online data mining system, is to mine perturbed data streams to cluster the data. Users can flexibly adjust the data attributes to be perturbed according to the security need. Therefore, threats and risks from releasing data can be effectively reduced.

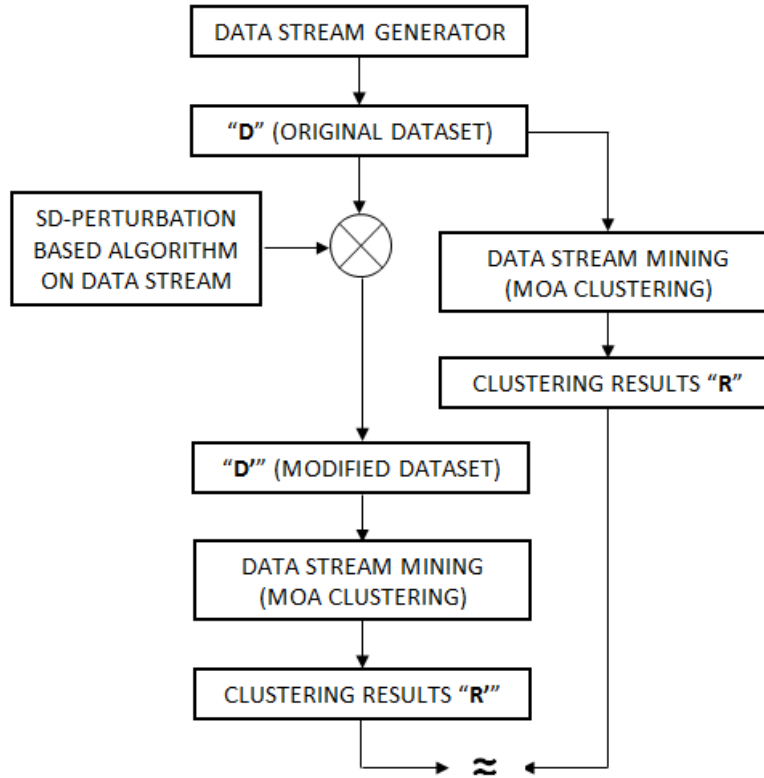


Figure 5.4: Extended framework of Massive Online Analysis (MOA) using SD-Perturbation: A Sensitive Drift based perturbation approach for privacy preserving in Data Stream Clustering

5.2.1 Formal Analysis of proposed work

The ultimate goal of data perturbation approach is to optimize the data transformation process by maximizing both, data privacy and data utility achieved. Proposed heuristic based privacy preserving in data stream mining using perturbation approach maintains the statistical properties of original data set for data mining process. Proposed approach not only perturbs the sensitive data alone but it also perturbs sensitive data using correlated tuple values, which means perturbed sensitive data is based on whole data set values. Purposed approach has no bias regard to sensitive data values during perturbation process. Proposed approach is not reversible. This means that adversary cannot retrieve the original data back from perturbed data set. Proposed approach maximizes the privacy on sensitive data and minimizes the information loss. Perturbation method attempts to preserve privacy of the data by modifying values of the sensitive attributes using SD-Perturbation approach which is value distortion based approach. Our proposed

approach work as follows:

- Step-1: On given original data set $D = I_1, I_2, I_3, I_4, I_5, \dots, I_n$, Calculate the Tuple T_i value $T = t_1, t_2, t_3, t_4, t_5, \dots, t_n$ for each instances from I_1 to I_n using following Equation except class attribute. Define the sensitive drift value SD and Sliding window size (W) in MOA.

$$T_i = Average \left\{ \sum_{i=1}^{W_n} - \frac{m}{Std_{dev}} \right\} \dots (1)$$

- Step-2: Based on our SD Perturbation approach, Sensitive value is modified using user defined sensitive drift value and tuple value as follow:
 - Step-2.1: Calculate User defined Sensitive drift (SD) (in percentage) value of each tuple value.

$$SD_{T_i} = \frac{T_i \times SD}{100} \dots (2)$$

- Calculate Upper bound (UB) and Lower bound (LB) for each tuple value basd on defined range (w) and select those tuple values which come in this range.

$$UB = T_i + SD_{T_i} \text{ and } LB = T_i - SD_{T_i} \dots (3)$$

- Based on Equation-3, if tuple value is in this range then select the corresponding Sensitive Values and calculate the average of these values which will give the final modified values of current instance.

$$SV'_i = Average(\sum_{i=1}^w SV_i) \dots (3)$$

Proposed approach maintaines the statistical properties such as mean, variance and standard deviation. Resultant of clustering (R') maintains the statistical properties with minimum information loss of original data stream.

Linear reconstruction of perturbed Data Stream D' to Original Data stream D is not possible because, generated perturbed streams D' using proposed method, the linear reconstruction is $D^* = D'R$, such that Discrepancy (D, D^*) is smaller one. If the perturbed streams D' and the original streams D are presented, the solution D^* can be simply recognized using linear regression. Since Data stream D is not available. Thus, in order to estimate D^* , certain extra constraints or assumptions must be imposed to make the problem solvable.

5.3 Proposed Algorithm

Proposed algorithm (Algorithm 2), perturbs sensitive attribute values before publishing for mining purpose. Proposed approach assumes only one sensitive attribute available in the data set and it can be further modified to extend the approach to perturb discrete set of attributes.

Algorithm 2 SD-Perturbation: A sensitive drift based perturbation approach for preserving privacy in data stream mining

Input: Data Stream $D = I1, I2, I3 \dots In$, Sensitive Drift SD , Window Size w

Intermediate Result: Perturbed Data stream $D' = I1, I2, I3 \dots In$

Output: Clustering Result set R and R' of Data Stream D and D'

Algorithm Steps

for Each Data set D **do**

Set $SA[i]$ ▷ store sensitive attribute values in array

end for

for each instance I in Data Set D **do** ▷ calculate the Tuple value using the average of normalized values except the class attribute

for $j=1$ to $n-1$ **do** ▷ n =number of attributes

if NOT (normalized (I_j)) **then**

$m = Mean(I_j)$

$v = Stdev(I_j)$

$z(a_j) = (value(I_j) - m) / v$

else

$z(a_j) = value(a_j)$

end if

end for

$Tvb[j] = AVG(I1toIn)$ ▷ store tuple values in arrays

end for

for $i=0$ to w **do**

$sp = 0$ ▷ sensitive percentage

$c_{iv} = Tvb[i]$	▷ Current Index Value
$sp = (c_{iv} * SD) / 100$	
$sv = 0$	▷ sliding value
for $j=I$ to $w+i$ do	
$rv = Tvb[j]$	▷ row value
$ub = rv + sp$	▷ Upper Bound
$lb = rv - sp$	▷ Lower Bound
if $(c_{iv} \leq ub \&\& c_{iv} \geq lb)$ then	
$sv = sv + SA[j]$	
end if	
end for	
$sv = AVG(sv)$	▷ Average of sv
$Value(IS) = sv$	▷ set the value of sensitive Instance
$Clustering(I)$	▷ Perform clustering process
end for	

5.4 Performance Evaluation

5.4.1 Experimental setup

To evaluate the effectiveness of proposed privacy preserving method, experiments have been carried out on Intel Core I3 Processor with 3 GB primary memory on Windows system. Simulation has been done in data stream clustering environment. We quantified proposed approach using resultant accuracy of true dataset clustering and perturbed dataset clustering. The experiments were processed on three datasets which are available from the UCI Machine Learning Repository and MOA dataset repository. K-Mean Clustering algorithm using WEKA data mining tool in MOA framework has been simulated to evaluate the accuracy of proposed PPDSM approach. MOA is a tool for implementing the methods and running experiments for online learning from evolving data streams (Bifet et al.). MOA supports evaluation of data stream learning algorithms on large streams for both Clustering and Classification. In addition to this, it also supports interface with WEKA machine learning algorithms. Following are the steps of using MOA frame-

work and showing how data stream mining with proposed perturbation technique works.

- Structure each dataset as streaming data in MOA framework.
- Define sliding window (W) over the data stream.
- Apply our proposed method to protect the sensitive attribute value to achieve privacy on dataset.
- Apply the K-mean method to find the clusters of perturbed dataset and original dataset. K-mean is scalable and known method which is used on static dataset and streaming data.
- Match the clusters of perturbed dataset with clusters of original dataset. F-measure is useful to measure the quality of clusters.

The proposed approach focussed on data perturbation through noise addition to preserving privacy of sensitive attributes. Tuple value based multiplicative data perturbation (TVB) (Chhinkaniwala and Garg) tried to retain statistical relationship between the attributes intact. Proposed approach considered private attribute as dependent attribute and remaining attributes of instance, other than class attribute, as independent attributes. Independent attributes of instances have been used to calculate instance specific random noise.

K-Mean clustering method over predefined sliding window size on perturbed data stream has been used in order to measure the accuracy and effectiveness of clustering outcomes over two standard data sets. Outcomes show that privacy has been achieved with more than 90% mining accuracy with test cases. Accuracy between original dataset and perturbed dataset has been quantified by percentage of instances assigned to different clusters with the help of cluster membership matrix. The proposed approach shows reasonably good results against evaluation measures - Precision, Recall, Misclassification and Cluster Membership Matrix (CMM). Experiments have been carried out to protect only numeric attributes.

5.4.2 Experimental Results

In this section, we focus on performance analysis of proposed method. The main benefit of our work over the existing one is that each party generates the perturbed data set based on sensitive drift. Experiments have been accomplished to measure accuracy and privacy while protecting sensitive data. For accuracy measurement, we have shown two different outcomes for analysis; 1) Represents clustering accuracy in terms of membership matrix which is derived from clustering result and 2) Represents graph for $F1_P$ (precision) and $F1_R$ (Recall) measures. For privacy Gain Measurement, here, we have calculated S (Security) using variance difference between original and Perturb dataset. Following Table 5.1 shows the data set used for test cases with attributes, which were perturbed by proposed algorithm. Nominal attributes have been ignored in proposed perturbation algorithm. Seven attributes from three standard data sets have been perturbed using proposed algorithm in the data stream scenario with sliding window size $w = 3000$ tuples.

Table 5.1: Description about datasets

Dataset	Input Instances	Nominal	Sensitive Attribute
Coverttype	65,000	Ignored	Elevation, Aspect, Slope
Electric Norm	45,000	Ignored	Nswprice, Nswdemand
Bank Marketing	45,000	Ignored	Balance, Duration

5.4.3 Cluster Membership Matrix (CMM)

In evaluation approach, we focussed on overall quality of generated clusters. We compared how closely each cluster in the perturbed dataset matches its corresponding cluster in the original dataset. The first need was to identify the matching of cluster by computing the matrix of frequencies. We refer to such a matrix as the Clustering Membership Matrix (CMM) shown in Table 5.2, where the rows represent the clusters in the original dataset, the columns represent the clusters in the perturbed dataset, and $Freq_{i,j}$ is the number of points in cluster C_i that falls in cluster C'_j in the perturbed dataset. After computing the frequencies $Freq_{i,j}$, we

Table 5.2: Clustering Membership Matrix (CMM)

	C1'	C2'		Cn'
C1	$Freq_{1,1}$	$Freq_{1,2}$	$Freq_{1,n}$
C2	$Freq_{2,1}$	$Freq_{2,2}$	$Freq_{2,n}$
:	:	:	:
Cn	$Freq_{n,1}$	$Freq_{n,2}$	$Freq_{n,n}$

scanned the CMM to calculate percentage of accuracy of perturbed data set for each cluster C'_i with respect to C_i in the original dataset.

In our experiment, we have compared the Clustering results of original dataset D to perturbed dataset D' where we have defined the sliding window size in Data stream $W=3000$. Sliding window size W means such number of records that are processed first in data stream. Once the process is over, sliding window moves to next. For clustering, we have used well-known K-mean Algorithm where we define $k=5$. We can also consider the user defined window size and sensitive drift in our experiments. For our experiment, we used user define window size 10, 30 and 50. According to sensitive drift (SD) and this window size, sensitive attribute value is perturbed. Figure 5.5 to 5.11 show the percentage of accuracy obtained for different sensitive attributes after performing clustering process on perturbed data set. Figure 5.5 to 5.7 show the accuracy obtained for Coverttype Dataset. Coverttype data set has three sensitive attributes named Elevation, Aspect and Slope. Accuracy is measured based on different window size (10, 30 and 50) and different sensitive drift values (3%, 5% and 10%). Based on these results, we have concluded that if sensitive drift is smaller then we achieve more information gain. We can also apply our algorithm on Electric norm and Bank marketing data stream and measure the accuracy which is shown in figure 5.8 to 5.11. Clustering accuracy of 90% (average) has been achieved over three standard datasets with sensitive drift values varying from 3% to 10%. From the graph, we can see that better accuracy is achieved if we minimize the window size. Based on our experiments we found that if window size is larger then, we will achieve less accuracy because in the process of perturbation, more instances are involved to modify the sensitive attribute values. In this graph, if we fix the window size to 10 then, we achieve better accu-

racy as compared to if we fix the window size to 30 or 50. In the graph W stands for user define window size and 'S' stands for user defined sensitive drift value in percentage.

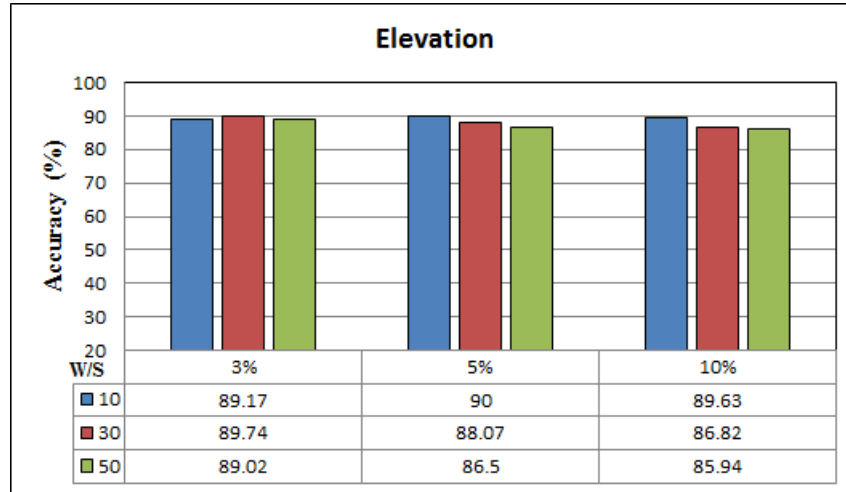


Figure 5.5: Accuracy obtained for Covertypes Dataset (Elevation)

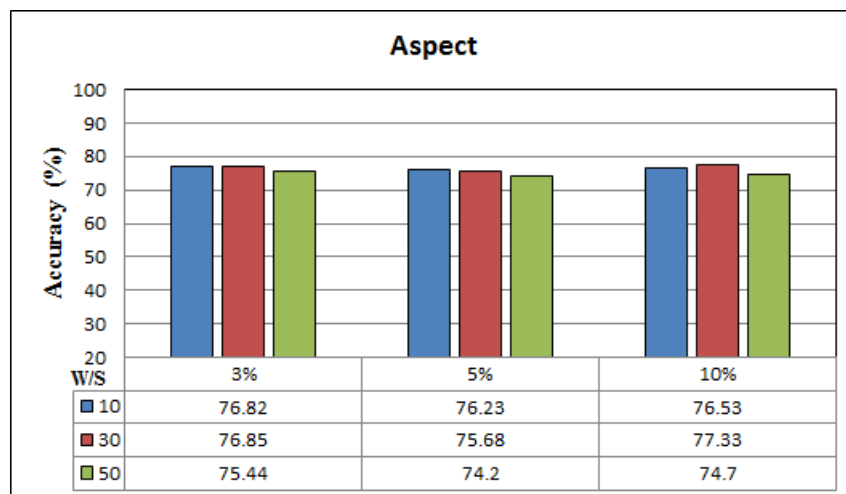


Figure 5.6: Accuracy obtained for Covertypes Dataset (Aspect)

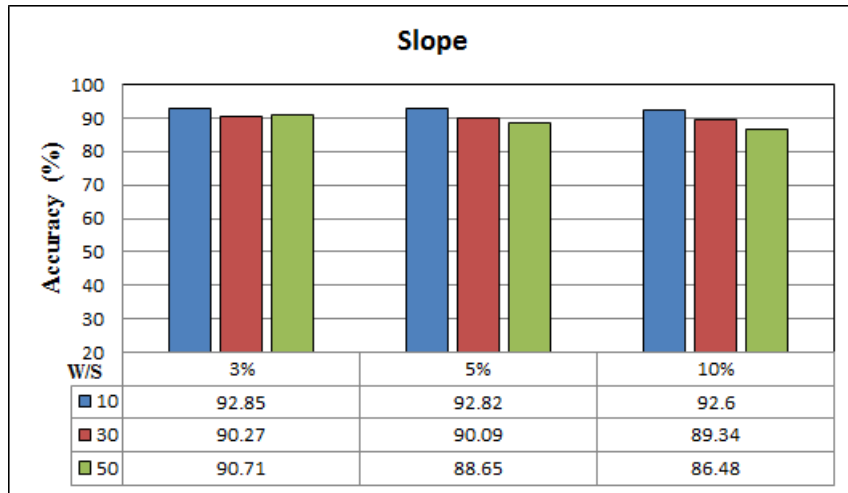


Figure 5.7: Accuracy obtained for Coverttype Dataset (Slope)



Figure 5.8: Accuracy obtained for ElectricNorm Dataset (swPrice)

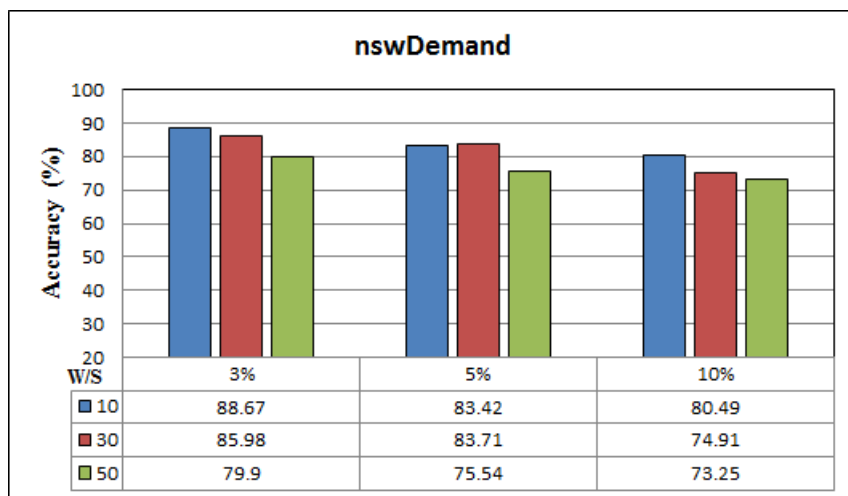


Figure 5.9: Accuracy obtained for ElectricNorm Dataset (swDemand)

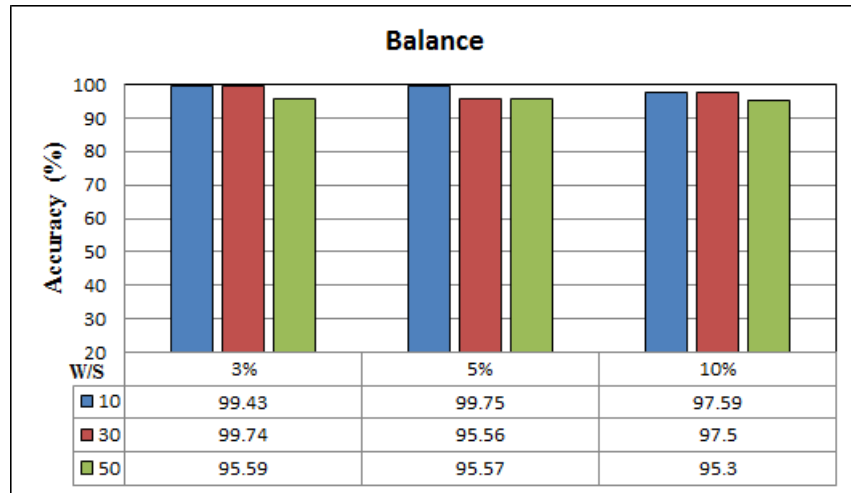


Figure 5.10: Accuracy obtained for Bank Marketing Dataset (Balance)

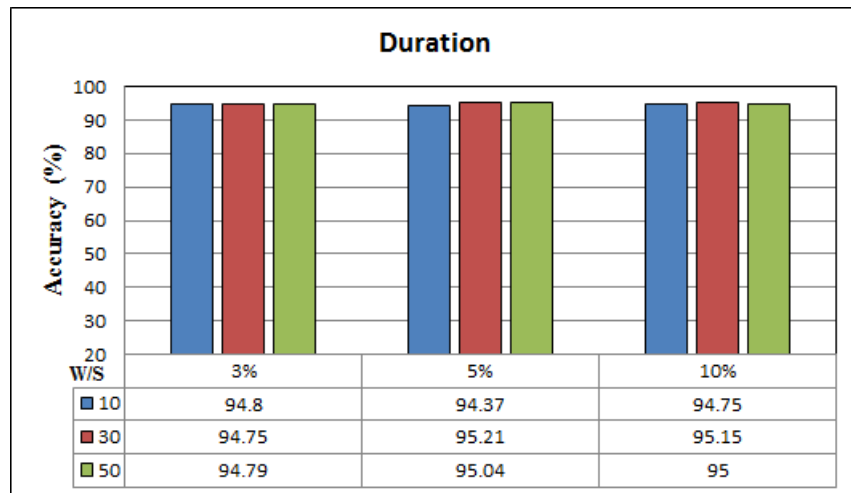


Figure 5.11: Accuracy obtained for Bank Marketing Dataset (Duration)

Table 5.3 shows the K-Means clustering result which shows that how many instances are correctly classified in each cluster and how many instances are not classified correctly means misclassify the instances. Table 5.4 indicates randomness in changes in mean and standard deviation values on three different attributes of standard data sets. Further, it shows no direct correlation between mean and standard deviation values of actual attribute values against perturbed values.

5.4.4 Precision & Recall Measurement

For Accuracy MOA framework provides different options. But, we mainly concentrated on two significant measures F1_P and F1_R. F1_P to define the precision

Table 5.3: K-Means cluster result of the perturbed datasets for window size=50 & Sensitive Drift=10%

Dataset	Attribute	Clusters					Miss Classifi- cation	Total	Accuracy (%)
		C1	C2	C3	C4	C5			
Bank	Balance	8731	9163	9596	8281	7116	2113	42887	95.30
Market	Duration	8707	9192	9455	8246	7152	2248	42752	95.00
Cover type	Aspect	12701	9134	7758	9291	10420	16696	49304	74.70
	Elevation	12883	12051	11224	10222	10338	9282	56718	85.94
	Slope	13185	12589	11018	9984	10303	8921	57079	86.48
Electric Norm	nsw De- mand	5231	8255	5594	8315	5569	12036	32964	73.25
	nswPrice	6245	8573	6811	10156	7453	5762	39238	87.20

of system by seeing the precision of individual cluster. F1 R defines the recall of system, which takes into account the recall of every single cluster. Outcomes are presented in terms of graphs for every single modified attribute and also for different range and sensitivity values. These are two significant measures to determine the usefulness and accuracy of the information retrieval system. Outcomes of proposed method have been quantified using precision and recall measures provided with MOA framework. Accuracy using these two measures is represented using line graph in figure 5.12, 5.13 and 5.14. Each graph represents the measure obtained when original data is processed without applying proposed perturbation approach and when data is undertaken through proposed approach. K-Means clustering algorithm is used to generate 5 clusters scenario. Precision and Recall measures have been evaluated with sliding window (w) 3000 tuples. Equations 1 and 2 represent the formula used to calculate precision and recall provided within MOA framework. Figure 5.12 to 5.14 show the result with different window size (w) and sensitive drift (s) values based on our proposed approach. It also shows the accuracy measurement between original and perturbed datasets using $f1_P$ and $f1_R$ measure. From this value we can conclude that accuracy of information gained after

Table 5.4: Parametric analysis on actual data values vs perturbed data values

Data set	Attribute	Actual		window	Sensitive Drift					
		Mean	StdDev		3%		5%		10%	
					Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Coverttype	Elevation	2359.3	254.89	10	627.51	71.89	627.55	71.04	627.57	68.91
				30	626.51	70.13	626.29	69.01	626.18	66.02
				50	625.93	69.03	625.58	67.82	625.44	64.67
ElectricNorm	Nswdemand	0.43	0.16	10	0.23	0.10	0.23	0.10	0.24	0.10
				30	0.23	0.10	0.23	0.09	0.24	0.08
				50	0.22	0.09	0.22	0.09	0.23	0.08
Bank Marketing	Balance	1426.7	3009.6	10	258.92	621.75	258.52	573.60	258.84	521.67
				30	256.46	517.58	255.96	463.60	256.20	405.25
				50	257.12	468.29	256.10	413.63	255.87	357.76

applying perturbation technique is preserved. Precision and Recall measures have almost the same values in all cases before and after data perturbation and hence it is proved that proposed approach provides better accuracy with minimum information loss and optimal gain in data privacy.

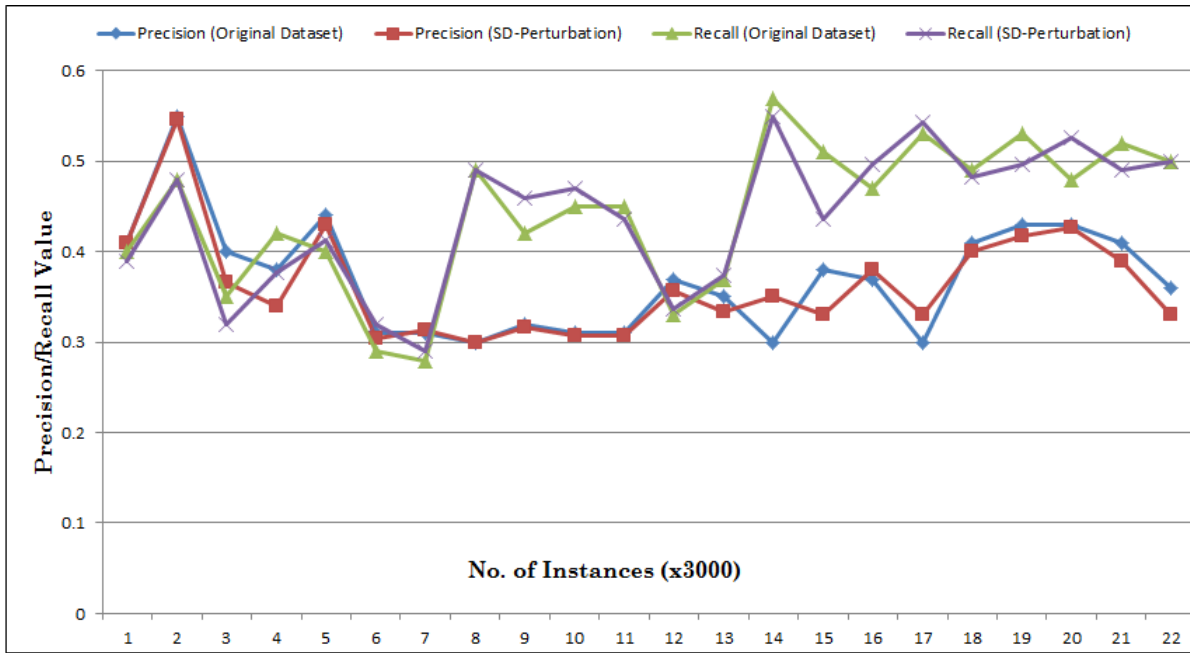


Figure 5.12: Accuracy measurement of Cover type dataset (Sensitive Attributes = (Aspect, Slope, Elevation) window size = 50, Sensitive drift = 10%)

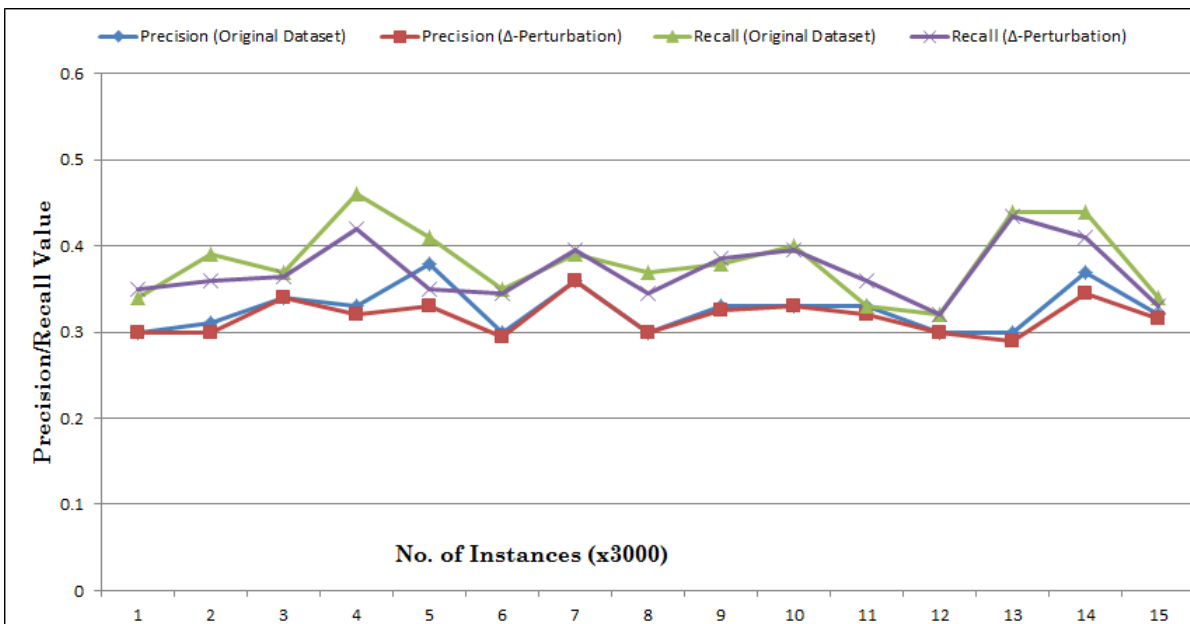


Figure 5.13: Accuracy measurement of Data Set Electric Norm (Sensitive Attributes = Demand, Price, window size=50, Sensitive drift=10%)

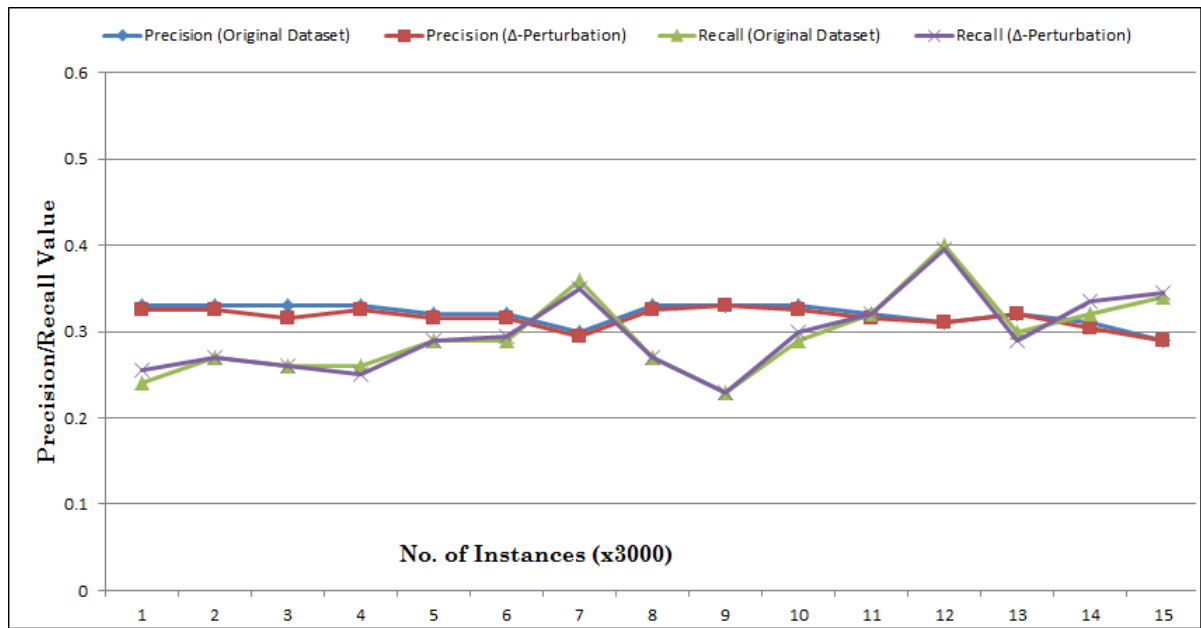


Figure 5.14: Accuracy measurement of Data Set Bank Marketing (Sensitive Attributes = Balance, Duration, window size=50, Sensitive drift=10%)

Table 5.5: Bias in Mean (BIM) and Bias in Standard Deviation (BISD)

Data sets	Attributes	Original Attributes vs. Perturbed Attributes					
		3% (SD)		5% (SD)		10% (SD)	
		BIM	BISD	BIM	BISD	BIM	BISD
Coverttype	Elevation	-0.73	-0.72	-0.73	-0.72	-0.73	-0.74
Electric norm	Nswdemand	-0.47	-0.4	-0.47	-0.42	-0.45	-0.46
Bank Marketing	Balance	-0.82	-0.82	-0.82	-0.84	-0.82	-0.86

Data error of the proposed algorithm is measured using BIM (Bias in Mean) and BISD (Bias in Standard Deviation) between original and perturbed values of sensitive attributes. Proposed algorithm has lower BIM and BISD values and hence there is no significant loss in data due to perturbation. The proposed method also show that there is no relation observed between values of mean and standard deviation computed for original data sets and perturbed data sets as shown in table.

5.4.5 Privacy Gain Measurement

Privacy gain is measured using the variance between real and modified data. we have measured the degree of privacy that is provided by SD-Perturbation: A drift

based perturbation approach for preserving privacy in data stream mining. This measure is given by $Var(X-X')$ where X represents an original data and X' represents the distorted data. This measure can be made scale invariant with respect to the variance of X by articulating security as:

$$S = \frac{Var(x-x')}{Var(x)}$$

As per this equation, the higher the value of 'S' (higher), then, higher the protection level. Figure 5.15 shows the Privacy Measurement using Variance difference in between original and perturbed dataset provided by these methods. After analyzing the graph, we can conclude that, variance difference in our proposed approach is higher than TVB (Tuple Value Based) and GDP (Geometric Data Perturbation) method. More variance difference means more security on sensitive attribute. Opponent can not reveal the original data back from perturbed data easily. Also, the information loss is less compared to other methods in our proposed approach.

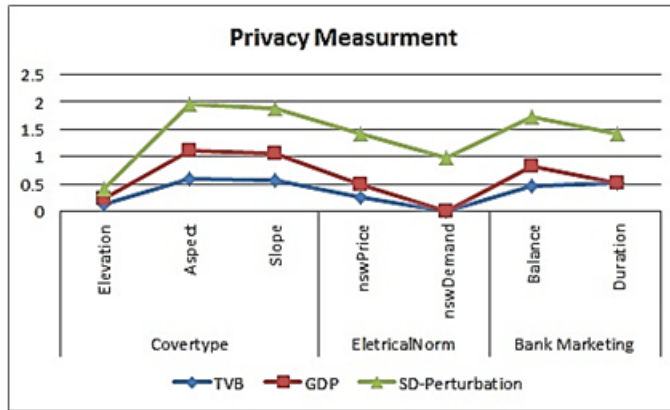


Figure 5.15: Privacy Measurement using Variance Difference

Another privacy gain is measured by PCA. PCA is applied on ElectroNorm Perturb dataset. Using PCA, find the coefficients for the principal components and variance of the respective principal components. Calculate the principal components of dataset D. Finally calculate the cumulative variance of principal components which is: 0.68, 0.98, 0.99, 1.0, 1.0, 1.0, 1.0, 1.0 and 1.0. In this case, the first principal component contains nearly 68% of the variance of the perturb dataset D. A data losing compression scheme, which discarded the second, third and fourth to nine principal components, would compress nine variables into 1, while losing only 32% of the variance. The other crucial thing to note about the principal

components is that they are completely uncorrelated. In this analysis, only approximate perturb dataset values will be recovered from perturbed data set but original data set is not recovered from perturbed data set. In terms of expose risk, given perturbation method given surety that third-party would not be able to get an accurate estimate of the value of a sensitive data.

5.5 Summary

Proposed research work tries to find out solutions to balance the data privacy and mining result. Several algorithms have been proposed, which understand the characteristics of the data set and perturb either sensitive attribute values or keep sensitive attribute's values unchanged and anonymized. SD-perturbation approach shows favorable results among other proposed approaches of data perturbation. Proposed approach uses the concept of the sensitive drift value, which is user-defined, so organization can decide how much privacy is needed. We applied K-Means clustering for the calculation of accuracy. Based on the characteristics of data set, we gained privacy and accuracy (we achieved, on an average 90% accuracy with 10% information loss after testing three different data sets) in almost all cases. The proposed method is flexible and is an easy-to-use method in the area of PPDSM like other existing methods. We quantified the privacy of our scheme using the concept of misclassification error. Information loss due to data perturbation was quantified by the loss of accuracy, which can be measured by percentage of instances of data stream which are mis-classified using cluster membership matrix. We also provided privacy level measurement which shows the level of security. In the next chapter, we have proposed another privacy preserving data stream mining algorithm which is applicable on multiple sensitive attributes.

Chapter 6

Heuristic based hybrid privacy preserving data stream mining approach using Perturbation and K-anonymization

In the previous chapter, we proposed SD -Perturbation: A Sensitive Drift based perturbation algorithm for privacy preserving in Data Stream Clustering. In this chapter, we have proposed Heuristic based hybrid privacy preserving data stream mining approach using Perturbation and K-anonymization.

Many data sets contain more than one sensitive attribute and it is necessary to make these sensitive data publicly available for analysis purpose. In spite of these, the issue of privacy needs to be addressed before streaming data is released for mining and analysis purposes.^{1, 2} With a view to addressing statistics of privateness worries, several techniques have emerged. Perturbation and K-anonymity has received significant attention over other privacy-preserving techniques be-

¹Part of this chapter has been published as Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Privacy Preserving in Data Stream Mining Using Multi-iterative K – Anonymization* in International Journal of Data Mining and Emerging Technologies, Volume 8, PP. 1-9, 2017

²Part of this chapter has been published as Paresh Solanki, Sanjay Garg, Hitesh Chhinkaniwala, *Heuristic-based hybrid privacy-preserving data stream mining approach using SD-Perturbation and Multi-Iterative K-Anonymization* in International Journal of Knowledge Engineering and Data Mining, Inderscience, Volume 5, No. 4, PP. 306-332, 2018

cause of its ease and effectiveness in guarding data. Our privacy-preserving hybrid approach is suitable for multiple sensitive attributes. To protect privacy, when a micro data set is distributed with several sensitive attributes for analysis, the data set should be anonymized (where anonymization is needed) and perturbed (where modification is needed) appropriately so that, the sensitive data of individuals cannot be recovered and hide the sensitive data with high superiority. Assume a data set that contains information about patients. Each record in that data set shows a patient by common attributes such as gender, age, address, marital status, job, salary, critical condition, etc. Such data sets can be used for medical study to find stimulating patterns by means of statistical analysis and data mining. Yet, the hospital is dedicated to secure the privacy of its patients and, subsequently, it cannot release the data set because the miner can associate this data set to other publicly available data sets and disclose the individual identity and know his/her critical condition and other important data. In other words, the table's characteristics that can be found in other openly accessible databases (known as the general population properties or quasi-identifier) possibly utilized by an enemy keeping in mind the end goal to take in account the particular qualities (known as the private traits-for a portion of the people) in the data set. We have to allow learning information about the public but not about the individuals who comprise such group. Many approaches were suggested for playing this gentle game that requires finding the precise route between data hiding and data disclosure.

The data owners have to convert records in such a way that if an opponent wants to find an individual's identity and to have understanding about QI (Quasi Identifiers), find out " $k-1$ " records that satisfy " $k-1$ " quasi identifiers and if an adversary wishes to locate the sensitive value, perturb records cannot disclose it. Data proprietors have to face such issues when several sensitive attributes exist in the records and which require special therapy using privacy preserving techniques. Suppose there is a data table having multiple sensitive attributes like condition and income. When a data proprietor focusses to guard one sensitive attribute, it may cause uncover of identity due to another one. So, we require a method to control all sensitive attributes. We can solve this problem if we use the combine strategy/multiple strategy (Privacy Preserving Techniques). K-anonymity

is preferred over the perturbative method because it does not compromise the integrity (truthfulness) of data. As a result, the anonymized data produced by the k-anonymity algorithm is reliable and useful for statistical analysis, research and data mining purposes. Data streams have a temporal dimension i.e. there is a maximum delay acceptable between inflowing data and its corresponding anonymized output. In some applications, the anonymized output triggers other actions. Hence, the receiving application should have strong guarantees on the maximum delay of its input data. In order to apply k-anonymity on data streams, current propositions in the literature incorporate a buffer/sliding window and delay constraints. The buffer holds the portion of the streaming data in order for the anonymisation process to occur. Delay constraints could be time or count-based. Time-based delay constraint specifies the time-duration for which a record can stay in the buffer while count-based specifies the maximum number of records the buffer can store. Figure 6.1 shows flow chart with combine strategies to preserve the privacy based on the data Perturbation and Anonymization methods. Due to combination of such types of varied techniques robust security can be obtained. Security result would be more effective in case of combined strategy rather than a single privacy preserving techniques used. We have proposed hybrid privacy preserving Algorithm to prevent multiple sensitive attributes using SD-Perturbation and Multi-iterative k-anonymization in data stream mining.

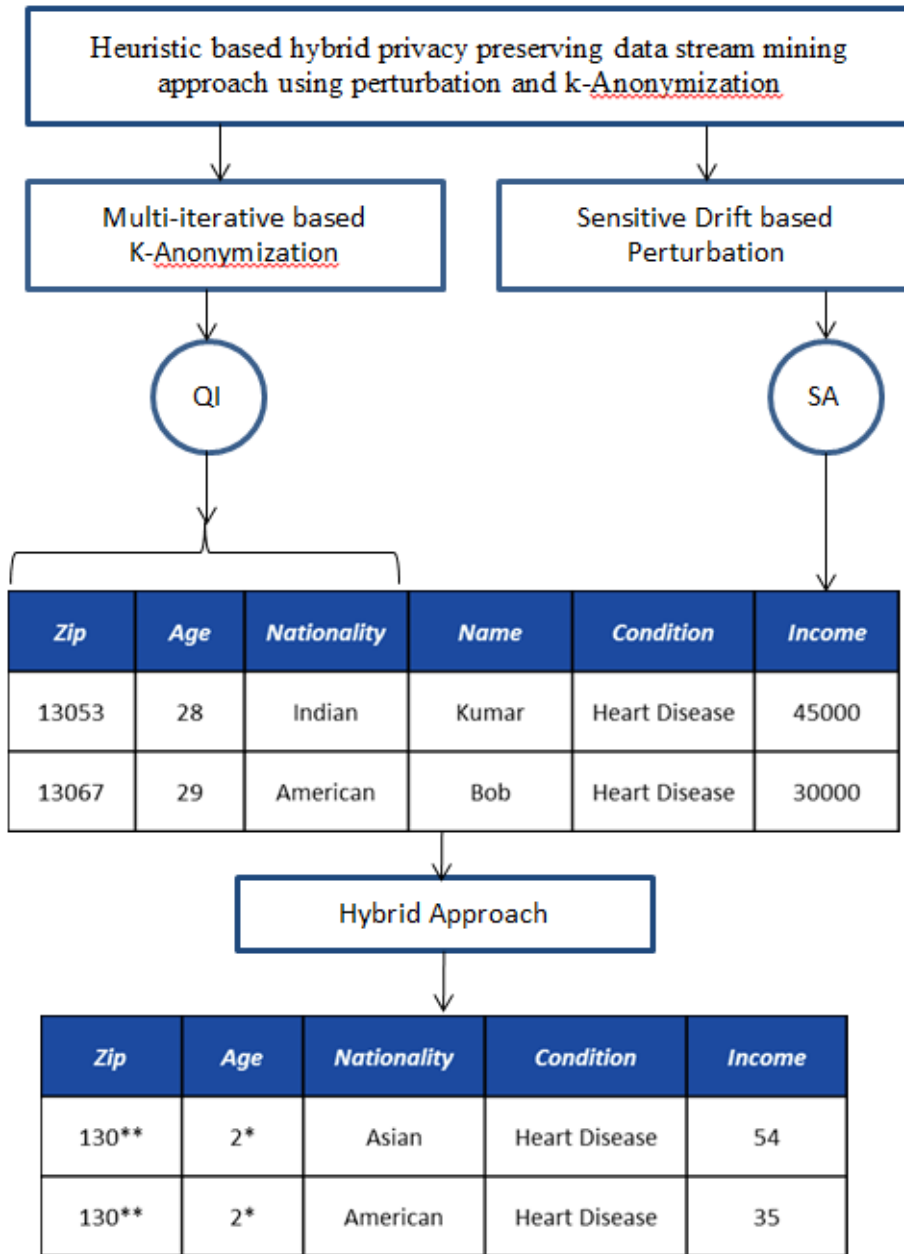


Figure 6.1: Flow chart of the proposed work

6.1 Problem Formulation

There are number of sectors like government sector or public sector where digital data is rising without bound at a high rate (e.g. millions) per day. Let $DStream$ be the data stream, t be the time stamp and w be the defined window size and m be the number of attributes into each data stream. Each data stream at different time stamps is represented as;

Data stream $DStream_1$: $DStream_1^{t1}, DStream_1^{t2}, DStream_1^{t3}, \dots, DStream_1^w$

Data stream DStream₂: $DStream_2^{t1}, DStream_2^{t2}, DStream_2^{t3}, \dots, DStream_2^w$

Data stream DStream₃: $DStream_3^{t1}, DStream_3^{t2}, DStream_3^{t3}, \dots, DStream_3^w$

.....

Data stream DStream_n: $DStream_n^{t1}, DStream_n^{t2}, DStream_n^{t3}, \dots, DStream_n^w$

Each data streams $DStream_{w \times m}$ can be mined using existing data stream mining algorithms.

6.2 Proposed framework

We have proposed a heuristic-based hybrid privacy-preserving data-mining approach using perturbation and K-anonymization to meet the following objectives: 1] Improving the response time. 2] Handling Privacy on Multiple Sensitive Attributes. 3] Minimizing Information loss vs. Maximizing Privacy protection. The objective of our proposed approach is to provide privacy before release of the data set. Original data set should be converted into perturbed and anonymized data set. Modified data set should generate identical result as of the original data set. Mathematically, we can represent our proposed approach as:

$$D'(Attributes, SA', QI') = D[Attributes, (SA + E), Anonymized(QI)]$$

Where, (D = Original Dataset, D' = Modified Dataset, SA = Original Sensitive Attribute, SA' = Modified Sensitive Attribute, QI = Quasi-identifiers, QI' = Modified Quasi-identifiers, E = Noise).

Protecting defendant's privacy while mining data or data stream mining generates challenges to data mining society. There are two methods how defendant's privacy can be preserved; one by altering sensitive information itself and second by keeping sensitive information as it is during the process of data mining but removing identifier attributes and generalizing and/or suppressing quasi-identifiers. A new term, called PRIVACYearn (Privacy earn or Privacy gain) has been coined while proposing a algorithm to generalize and/or suppress quasi-identifiers value. Selective anonymization has been applied based on PRIVACYearn computed in subsequent iteration.

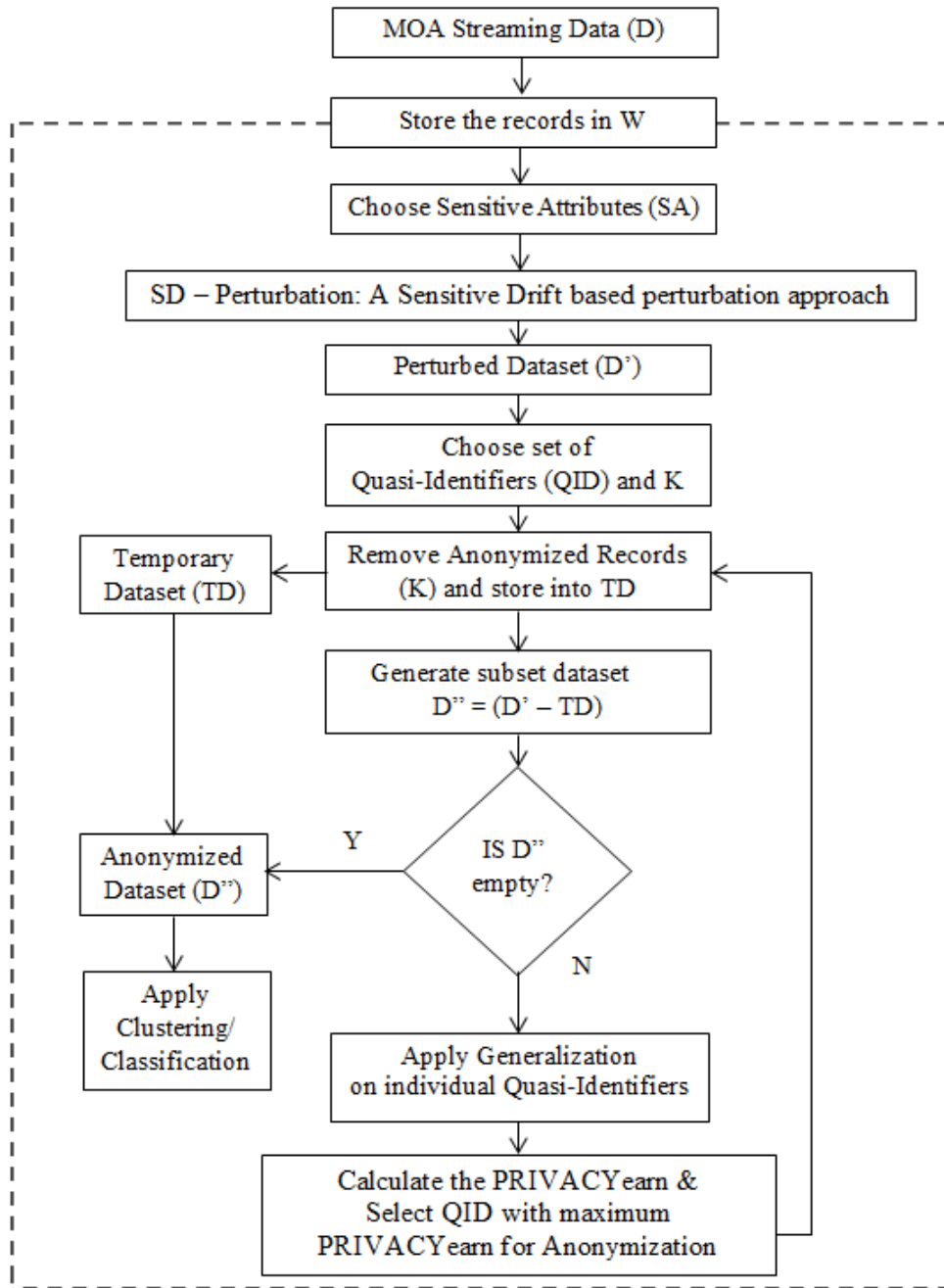


Figure 6.2: Framework of Heuristic based hybrid privacy preserving data mining approach using Perturbation and K-anonymization

In proposed hybrid algorithm, SD-Perturbation based privacy preserving data stream mining approach is applied on sensitive attribute values and Multi-Iterative K-Anonymization is applied on quasi-identifiers. PRIVACYearn has been computed for individual quasi-identifiers upon generalization based on dimension table. Quasi-identifier with maximum PRIVACYearn has been selected and generalized instances have been marked k-anonymized. Anonymized instances are then

removed from data set D' and stored into temporary data set TD. For rest of the unanonymized instances in D' , the same process is applied until k-anonymization is achieved for entire data set D' . Finally, store temporary data set TD into D'' . Proposed algorithm has an advantage of rescanning of only subset of original data set which is not k-anonymized and algorithm works very well on streaming data based on sliding window size. We can achieve the privacy on data as per user define window size. Proposed PRIVACYearn based algorithm has been tested against standard data sets and it has been found that defendant's privacy is protected with small amount of data loss which occurs due to anonymization. Classification results with original data sets versus classification results obtained using anonymized data sets over Naive Bayes algorithm have been compared. It has been observed that the proposed approach significantly reduces execution time because total data set rescan is no more required and provides better privacy without compromising much on data utility. Figure 6.2 shows the framework of proposed approach.

6.3 Proposed Algorithm

The proposed heuristic-based hybrid PPDM approach is the combination of two methods: Perturbation and K-anonymization. Perturbation is applied on sensitive attributes and k-anonymization, where we applied the generalization on quasi-identifiers. In this approach, we focus more on the utilization of information with minimum loss. In this approach, we will check the privacy gain after performing hybrid approach on the given dataset. It can be simply demonstrated that, snooper will get the original data back easily, if we apply privacy-preserving techniques alone. For this problem, instead of applying only one method, we applied two or more methods combinedly, to accomplish the goal. The proposed hybrid approach for privacy preserving in data stream mining approach maintains the statistical properties of original data set for data-mining process. The approach has no bias regarding sensitive data values during perturbation process. The proposed approach is not reversible which means that the adversary cannot retrieve the real data back from the modified (perturbed) data set. The proposed approach maximizes the privacy on sensitive data and minimizes the information loss. In

database community, researchers have designed numerous methods that execute records in a group-based fashion, which transform the records in a way that preserves precise privacy metrics. Such altered records can be distributed without fear of restoration through attacks. There is a supposition that certain attributes of a data set encompass quasi-identifiers that distinctively recognize individual records which are related, as well as sensitive attributes that must not be connected to the individual by an untrusted opponent.

Algorithm 3 Heuristic based hybrid privacy preserving data mining approach using Perturbation and K-anonymization

Input: Streaming Dataset D , Sensitive Drift SD , Number of Quasi-Identifiers QI , Sensitive attribute SA , Value of K , sliding window size W .

Intermediate Result: Perturbed Data stream D' and anonymized dataset D''

Output: Clustering and classification Result set R and R' of Data Stream D and D'

Algorithm Steps

W = Sliding window size

SD = Sensitive Drift

D' = Perturbed Dataset

K = Value of K

QI = Quasi-identifiers

D = contain R Records

$R1$ = Un-Anonymized Records

$R2$ = Anonymized Records

D'' = Final anonymized Records

for each instance of D **do**

for $i=1$ to W **do**

 Call SD-Perturbation Algorithm \triangleright Perform Data Perturbation using Sensitive Drift

 Store the Perturbed data into D

if ($\text{Size}(W_i) < K$) **then**

 Anonymization is not possible

 Return

end if

for $j=1$ to W_i **do**

for each quasi identifier QI **do**

$R2$ = store already anonymized data

$R1 = W_j - R2$ (Un-Anonymized Records)

$R2$ = store anonymized data after generalization

 Uniquely apply generalization on $R1$

$$Privacy = (R2 * 100) / W_i$$

end for

end for

end for

end for

Repeat the step above until we are getting maximum Privacy with changing the generalization level

Privacy = Max(Privacy)

D = R2 Do the clustering/classification on D

How the proposed multi-iterative k-anonymization works is explained through the following example. In given medical dataset 6.3, there are two sensitive attributes: "condition" and "salary".

Non-sensitive data				Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	28	Male	Heart Disease	20000
2	13068	29	Male	Heart Disease	35000
3	13068	21	Female	Viral Infection	15000
4	13053	23	Male	Viral Infection	5000
5	14853	50	Female	Cancer	50000
6	14853	55	Male	Heart Disease	25000
7	14850	47	Female	Viral Infection	5000
8	14850	49	Female	Viral Infection	15000
9	13053	31	Male	Cancer	35000
10	13053	37	Male	Cancer	40000
11	13068	36	Female	Cancer	50000
12	13068	35	Female	Viral Infection	5000
13	14850	32	Male	Heart Disease	20000
14	13053	42	Male	Hepatitis	15000
15	13068	40	Female	Brochitis	5000
16	14850	28	Male	Broken Arm	15000
17	13053	40	Male	Viral Infection	35000
18	13053	25	Female	Flu	20000
19	14853	51	Female	Cancer	10000
20	13068	55	Male	Heart Disease	25000

Anonymization:	k=3		
Quasi-Identifiers	ZIP,AGE,GENDER		
Generalization Levels:	Level-1	Level-2	Level-3
ZIP:	*	**	***
AGE:	Interval	Young	
		Mid Age	
		Older	
GENDER	Person		

Figure 6.3: Medical Dataset

Step-1: Remove records from original dataset which are already Anonymized. Anonymization process will start from single quasi-identifier to multiple quasi-identifiers and will calculate the PRIVACY_{yearn} based on following Equation.

$$PRIVACY_{\text{Year}} = (\text{Number of Anonymized Records} * 100) / \text{Total Number of Records}$$

Step-2: Select the "Gender" as Quasi-Identifier and apply the Generalization as Level-1 and calculate the PRIVACY_{Year}. The resultant of PRIVACY_{Year} is 0%. Next, select the Pincode as Quasi-Identifier and apply the Generalization as Level-1 and calculate the PRIVACY_{Year}. The resultant of PRIVACY_{Year} is 0%. Next, Select the Age as Quasi-Identifier and apply the Generalization as Level-1 and calculate the PRIVACY_{Year}, which is 15%. Next, Remove the anonymized records and store them into another repository. Select the two Quasi-Identifiers and apply the generalization level as per given dimension table. Same Process is repeated. Finally, selects the appropriate Quasi-Identifiers which gives best PRIVACY_{Year}. The entire process is reflected in shown in 6.4 to 6.13.

	Non-sensitive data			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	28	Person	Heart Disease	20000
2	13068	29	Person	Heart Disease	35000
3	13068	21	Person	Viral Infection	15000
4	13053	23	Person	Viral Infection	5000
5	14853	50	Person	Cancer	50000
6	14853	55	Person	Heart Disease	25000
7	14850	47	Person	Viral Infection	5000
8	14850	49	Person	Viral Infection	15000
9	13053	31	Person	Cancer	35000
10	13053	37	Person	Cancer	40000
11	13068	36	Person	Cancer	50000
12	13068	35	Person	Viral Infection	5000
13	14850	32	Person	Heart Disease	20000
14	13053	42	Person	Hepatitis	15000
15	13068	40	Person	Brochitis	5000
16	14850	28	Person	Broken Arm	15000
17	13053	40	Person	Viral Infection	35000
18	13053	25	Person	Flu	20000
19	14853	51	Person	Cancer	10000
20	13068	55	Person	Heart Disease	25000

Figure 6.4: Anonymized Dataset ($Gender_1$), Privacy Gain=0%

	Non-sensitive data			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	1305*	28	Male	Heart Disease	20000
2	1306*	29	Male	Heart Disease	35000
3	1306*	21	Female	Viral Infection	15000
4	1305*	23	Male	Viral Infection	5000
5	1485*	50	Female	Cancer	50000
6	1485*	55	Male	Heart Disease	25000
7	1485*	47	Female	Viral Infection	5000
8	1485*	49	Female	Viral Infection	15000
9	1305*	31	Male	Cancer	35000
10	1305*	37	Male	Cancer	40000
11	1306*	36	Female	Cancer	50000
12	1306*	35	Female	Viral Infection	5000
13	1485*	32	Male	Heart Disease	20000
14	1305*	42	Male	Hepatitis	15000
15	1306*	40	Female	Brochitis	5000
16	1485*	28	Male	Broken Arm	15000
17	1305*	40	Male	Viral Infection	35000
18	1305*	25	Female	Flu	20000
19	1485*	51	Female	Cancer	10000
20	1306*	55	Male	Heart Disease	25000

Figure 6.5: Anonymized Dataset ($Zipcode_1$), Privacy Gain=0%

	Non-sensitive data			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Male	Heart Disease	20000
2	13068	21-30	Male	Heart Disease	35000
3	13068	21-30	Female	Viral Infection	15000
4	13053	21-30	Male	Viral Infection	5000
5	14853	41-50	Female	Cancer	50000
6	14853	51-60	Male	Heart Disease	25000
7	14850	41-50	Female	Viral Infection	5000
8	14850	41-50	Female	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Female	Cancer	50000
12	13068	31-40	Female	Viral Infection	5000
13	14850	31-40	Male	Heart Disease	20000
14	13053	41-50	Male	Hepatitis	15000
15	13068	31-40	Female	Brochitis	5000
16	14850	21-30	Male	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Female	Flu	20000
19	14853	51-60	Female	Cancer	10000
20	13068	51-60	Male	Heart Disease	25000

Figure 6.6: Anonymized Dataset (Age_1), Privacy Gain=15%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	13068	21-30	Person	Heart Disease	35000
3	13068	21-30	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	14853	41-50	Person	Cancer	50000
6	14853	51-60	Person	Heart Disease	25000
7	14850	41-50	Person	Viral Infection	5000
8	14850	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	14850	31-40	Person	Heart Disease	20000
14	13053	41-50	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	14850	21-30	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	14853	51-60	Person	Cancer	10000
20	13068	51-60	Person	Heart Disease	25000

Figure 6.7: Anonymized Dataset ($Age_1, Gender_1$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	13068	young	Person	Heart Disease	35000
3	13068	young	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	14853	mid Age	Person	Cancer	50000
6	14853	older	Person	Heart Disease	25000
7	14850	mid Age	Person	Viral Infection	5000
8	14850	mid Age	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	14850	mid Age	Person	Heart Disease	20000
14	13053	mid Age	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	14850	young	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	14853	older	Person	Cancer	10000
20	13068	older	Person	Heart Disease	25000

Figure 6.8: Anonymized Dataset ($Age_2, Gender_1$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	1306*	21-30	Person	Heart Disease	35000
3	1306*	21-30	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	1485*	41-50	Person	Cancer	50000
6	1485*	51-60	Person	Heart Disease	25000
7	1485*	41-50	Person	Viral Infection	5000
8	1485*	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	1485*	31-40	Person	Heart Disease	20000
14	1305*	41-50	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	1485*	21-30	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	1485*	51-60	Person	Cancer	10000
20	1306*	51-60	Person	Heart Disease	25000

Figure 6.9: Anonymized Dataset ($Age_1, Gender_1, Zipcode_1$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	1306*	young	Person	Heart Disease	35000
3	1306*	young	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	1485*	41-50	Person	Cancer	50000
6	1485*	older	Person	Heart Disease	25000
7	1485*	41-50	Person	Viral Infection	5000
8	1485*	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	1485*	mid age	Person	Heart Disease	20000
14	1305*	mid age	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	1485*	young	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	1485*	older	Person	Cancer	10000
20	1306*	older	Person	Heart Disease	25000

Figure 6.10: Anonymized Dataset ($Age_2, Gender_1, Zipcode_1$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	130**	21-30	Person	Heart Disease	35000
3	130**	21-30	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	1485*	41-50	Person	Cancer	50000
6	148**	51-60	Person	Heart Disease	25000
7	1485*	41-50	Person	Viral Infection	5000
8	1485*	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	148**	31-40	Person	Heart Disease	20000
14	130**	41-50	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	148**	21-30	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	148**	51-60	Person	Cancer	10000
20	130**	51-60	Person	Heart Disease	25000

Figure 6.11: Anonymized Dataset ($Age_1, Gender_1, Zipcode_2$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	13***	21-30	Person	Heart Disease	35000
3	13***	21-30	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	1485*	41-50	Person	Cancer	50000
6	14***	51-60	Person	Heart Disease	25000
7	1485*	41-50	Person	Viral Infection	5000
8	1485*	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	14***	31-40	Person	Heart Disease	20000
14	13***	41-50	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	14***	21-30	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	14***	51-60	Person	Cancer	10000
20	13***	51-60	Person	Heart Disease	25000

Figure 6.12: Anonymized Dataset ($Age_1, Gender_1, Zipcode_3$), Privacy Gain=60%

	Non-sensitive Attribute			Sensitive Attribute	
	ZIP	AGE	GENDER	CONDITION	SALARY
1	13053	21-30	Person	Heart Disease	20000
2	1****	young	Person	Heart Disease	35000
3	1****	young	Person	Viral Infection	15000
4	13053	21-30	Person	Viral Infection	5000
5	1485*	41-50	Person	Cancer	50000
6	1****	older	Person	Heart Disease	25000
7	1485*	41-50	Person	Viral Infection	5000
8	1485*	41-50	Person	Viral Infection	15000
9	13053	31-40	Male	Cancer	35000
10	13053	31-40	Male	Cancer	40000
11	13068	31-40	Person	Cancer	50000
12	13068	31-40	Person	Viral Infection	5000
13	1****	mid age	Person	Heart Disease	20000
14	1****	mid age	Person	Hepatitis	15000
15	13068	31-40	Person	Brochitis	5000
16	1****	young	Person	Broken Arm	15000
17	13053	31-40	Male	Viral Infection	35000
18	13053	21-30	Person	Flu	20000
19	1****	older	Person	Cancer	10000
20	1****	older	Person	Heart Disease	25000

Figure 6.13: Anonymized Dataset ($Age_2, Gender_1, Zipcode_4$), Privacy Gain=90%

6.4 Performance Evaluation

Simulation has been done in data stream clustering environment. We quantified proposed approach using resultant accuracy of true dataset clustering and anonymized with perturbed dataset clustering. The experiments were processed on two different data sets (Jock A. Blackard, Kohavi and Becker). Proposed algorithm for data anonymization has been developed using Java. Classification results have been compared over Naive Bayes algorithm available within MOA tool. Table 6.3 and 6.2 show the quasi-identifiers with generalization levels. Figure 6.14 shows the Performance of accuracy using Multi-iterative k-Anonymization algorithm. It has been observed that while applying proposed algorithm based on Privacy gain, defendant's information is protected with minimal data loss. Further this approach is based on multi-iteration on non k-anonymized set of tuples that greatly reduces processing time. Earlier proposed works were rescanning entire data set for anonymization, while we have made an attempt to provide solution with lesser execution time by proposing Privacy Gain concept. Table 6.3 shows

the Privacy gain aggregate result (all sliding window). Generalization has been applied for each quasi-identifier. Privacy gain has been computed for individual quasi-identifiers upon generalization based on dimension table, which is shown in Table 6.1 and 6.2. Quasi-identifier with maximum privacy gain has been selected and tuples generalized have been marked k- anonymized. Anonymized tuples are then removed from data set D' and stored into temporary table TD. For the rest of unanonymized tuples in D' , the same process has been applied until k-anonymization is achieved for entire data set DS or generalization is applied to all quasi-identifiers to the maximum level as mentioned in dimension table and no further generalization is possible on unanonymized tuples. Table 6.4 shows the Privacy gain aggregate result with execution time. Proposed algorithm is integrated in MOA framework. Bank marketing and adult data sets have been analyzed with different set of quasi-identifiers and levels of generalization using sliding window concept.

Table 6.1: Quasi-identifiers in Bank Marketing data set with applied generalization level for k-anonymization

Quasi-identifiers	Age, Job, marital-status		
Generalization Level	Level-0	Level-1	Level-2
Age	Age	*	**
Job	management	commercial	known
	technician	technical	known
	entrepreneur	commercial	known
	blue-collar	commercial	known
	unknown	not known	unkown
	retired	not known	unkown
	admin.	commercial	known
	services	not known	unkown
	self-employed	not known	unkown
	unemployed	not known	unkown
	housemaid	not known	unkown
	student	study	known
Marital-Status	married	married	married
	single	single	single
	divorced	married but single	single

Table 6.2: Quasi-identifiers in Adult data set with applied generalization level for k-anonymization

Quasi-identifiers	Age, Occupation, Sex, Relationship, Native Country, Race, Workclass		
Generalization Level	Level-0	Level-1	Level-2
Age	Age	*	**

Occupation	Adm-clerical	class3	low
	Exec-managerial	class1	high
	Handlers-cleaners	class4	low
	Prof-specialty	class1	high
	Other-service	class4	low
	Sales	class3	low
	Craft-repair	class3	low
	Transport-moving	class4	low
	Farming-fishing	class4	low
	Machine-op-inspct	class3	low
	Tech-support	class2	high
	Protective-serv	class2	high
	Armed-Forces	class2	high
	Priv-house-serv	class4	low
Sex	Male / Female	Person	
Relationship	Not-in-family	Not-in-family	
	Husband	In family	
	Wife	In family	
	Own-child	In family	
	Unmarried	In family	
	Other-relative	In family	
Native Country	Country names	Asia	
		Europe	Asia, Middle East & Gulf-group1
		Middle East and Gulf	Americas & Europe-group2
		Americas	
Race	White	ame	
	Black	asi	

	Asian-Pac-Islander	asi	
	Amer-Indian-Eskimo	ame	
	Other		
Workclass	State-gov	government	government
	Self-emp-not-inc	own business	private
	Private	private	private
	Federal-gov	government	government
	Local-gov	government	government
	?	?	?
	Self-emp-inc	own business	private
	Without-pay	private	private
	Never-worked	not known	private

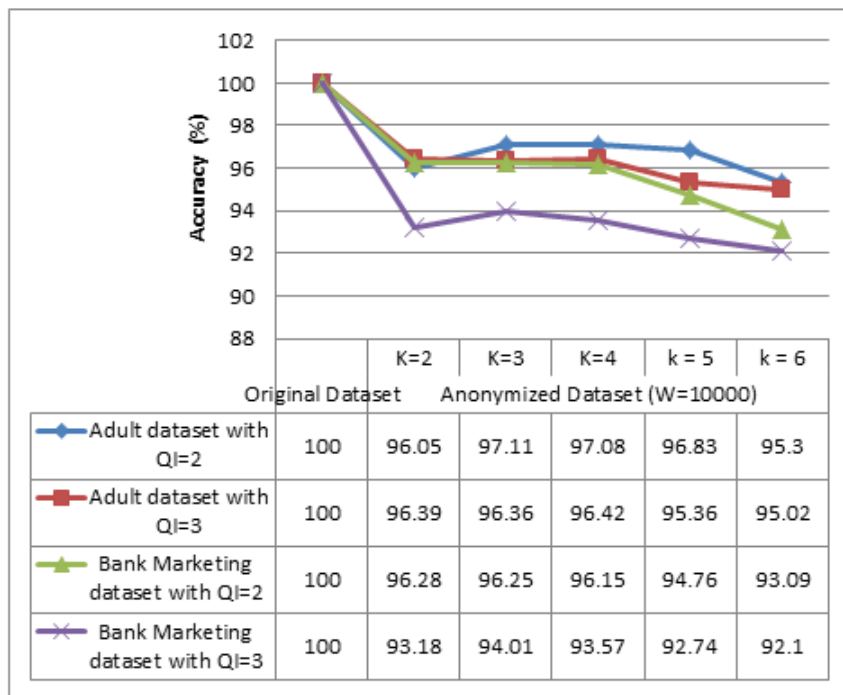


Figure 6.14: Accuracy measured using Naive Bayes algorithm on Adult and Bank Marketing Dataset

Table 6.3 assumes 3 quasi-identifiers with generalization levels as indicated. 3-anonymization has been applied on Bank marketing data set. As seen from Table 6.3, Adult data set has default 3-anonymization. 44.37% of 3-anonymization is

Table 6.3: Privacy Gain result of Bank Marketing Data set (k=3, w=3000)

level	Quasi-identifier Generalization	Overall Performance	
		Privacy Gain (%)	Execution Time (min)
0		1.16	0:00:00
1	age-1	14.29	2:08:51
2	age-2	25.03	3:06:42
3	job-1	34.23	3:55:43
4	job-2	41.61	4:39:41
5	marital-1	44.37	5:08:29

possible with proposed algorithm but 41.61% privacy is the best tradeoff between 3-anonymization and execution time. Table 6.3 show the overall performance (Privacy gain and Execution time) for different value of k. Privacy gain is decreased when value of k is increased.

Table 6.4: Privacy Gain result of Adult and Bank Marketing Dataset (W=10000)

Dataset	Quasi-identifier Generalization	value of k	Overall Performance	
			Privacy Outcome (%)	Execution Time (min)
Bank Marketing	AGE-2, JOB-2, MARITAL-1	3	44.37	5:08:29
		4	37.97	5:17:13
		5	32.98	5:23:29
		6	28.92	3:27:45
Adult	AGE-2, OCCUPATION-2, SEX-1, RELATIONSHIP-1, NATIVE COUNTRY-2, RACE-1, WORKCLASS-2	3	97.69	3:02:18
		4	97.16	3:03:20
		5	96.73	3:50:23
		6	96.21	5:45:55

Following figure 6.15 and 6.16 show the privacy outcome in percentage and logarithmic trends according to value of k.

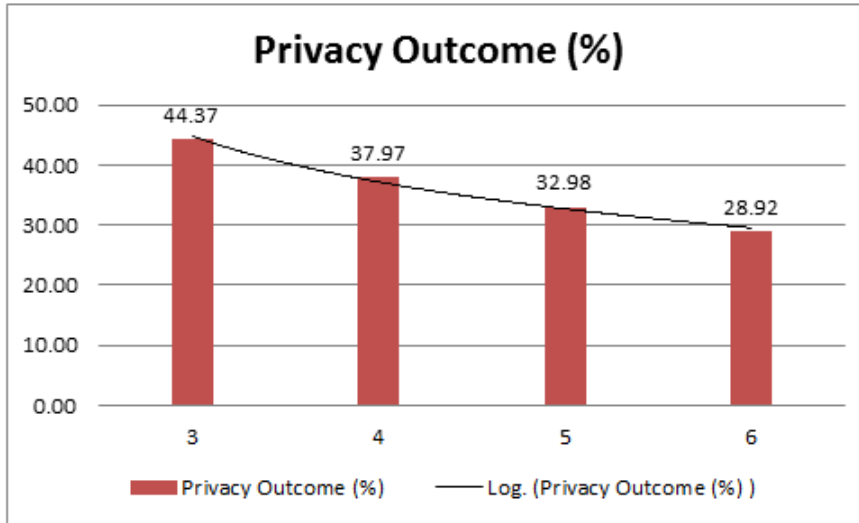


Figure 6.15: Privacy Gain of bank marketing data set

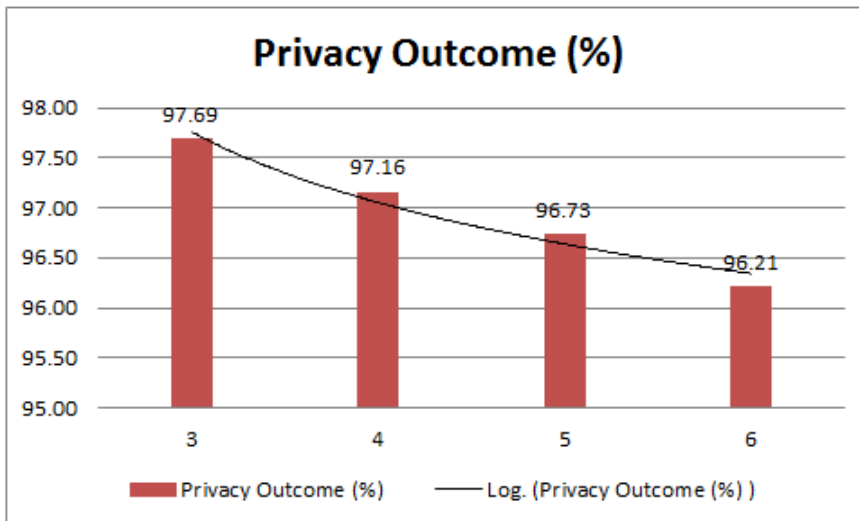


Figure 6.16: Privacy Gain of Adult data set

Adult data set and Bank marketing data set have been anonymized with different set of quasi-identifiers, generalization levels and anonymization. Modified data sets have been classified using Naive Bayes classifiers and the results show that average 85% classification accuracy is achieved after anonymizing dataset. Figure 6.17 and 6.18 show the privacy gain results of bank and adult dataset at different window sizes. Table 6.5 and 6.6 show the Privacy gain based on multi-iterative k-Anonymization with window size 3000 and 10000. Table 6.5 and 6.6 show the Privacy gain result with different values of K and Quasi-identifiers.

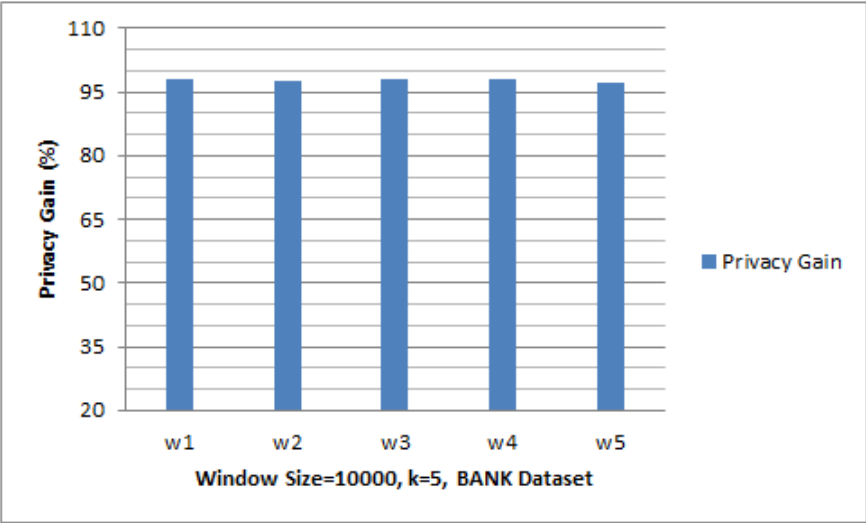


Figure 6.17: Privacy gain outcome of bank dataset (window size=10000)

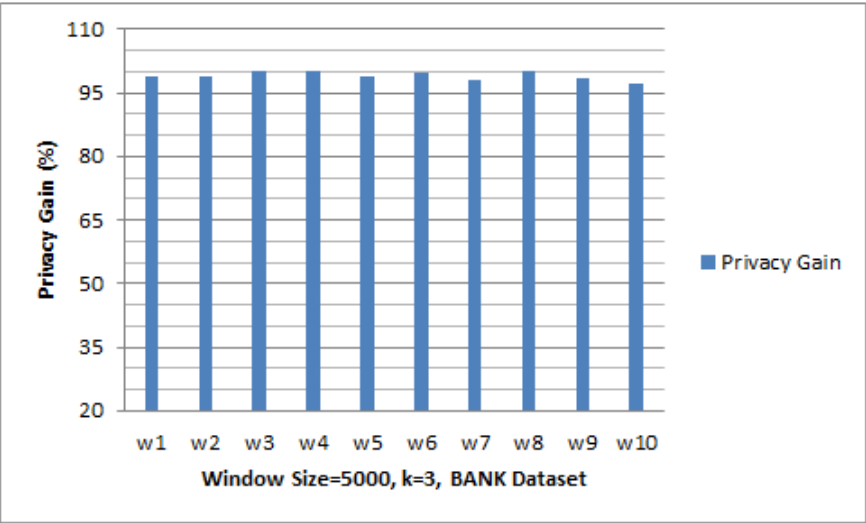


Figure 6.18: Privacy gain outcome of bank dataset (window size=5000)

Table 6.5: Performance of PRIVACYearn based Multi-iterative k-Anonymization algorithm (window size=3000)

Dataset	Records processed	Correctly classified using Naive Bayes algorithm (%)					
		Original Dataset	Anonymized Dataset (QI=2)				
			K=2	K=3	K=4	k = 5	k = 6
Adult	48000	100	95.14	95.06	96.76	95.81	94.27
Bank Marketing	45000	100	94.26	94.12	94.34	92.16	93.28
			Anonymized Dataset (QI=3)				
Adult	48000	100	95.92	94.16	94.02	94.76	94.23
Bank Marketing	45000	100	92.15	92.19	92.52	92.74	91.05

Table 6.6: Performance of PRIVACYearn based Multi-iterative k-Anonymization algorithm (window size=10000)

Dataset	Records processed	Correctly classified using Naive Bayes algorithm (%)					
		Original Dataset	Anonymized Dataset (QI=2)				
			K=2	K=3	K=4	k = 5	k = 6
Adult	48000	100	96.05	97.11	97.08	96.83	95.30
Bank Marketing	45000	100	96.28	96.25	96.15	94.76	93.09
			Anonymized Dataset (QI=3)				
Adult	48000	100	96.39	96.36	96.42	95.36	95.02
Bank Marketing	45000	100	93.18	94.01	93.57	92.74	92.10

6.4.1 Comparison of proposed approach with K-Anonymization

Table 6.7 and 6.8 shows the comparison of proposed approach with existing k-anonymization (Sweeney). We have measured the privacy outcomes and execution time for different values of "K". Result shows that, proposed approach is taking a less execution time in comparison with existing approach for different values of "K". Figure 6.19 shows that, privacy outcomes are better than existing approach.

Table 6.7: Comparison of proposed approach with existing k-anonymization approach (Privacy outcomes)

Dataset	Quasi-identifier Generalization	value of k	Multi-Iterative K-Anonymization based	K-Anonymization based
			Privacy Outcome(%)	Privacy Outcome (%)
Bank Marketing	<Age-2,Job-2,Marital-1>	3	44.37	43.3
		4	37.97	36.8
		5	32.98	32.8
		6	28.92	28.89
Adult	<Age-2,Occupation-2, Sex-1, Relationship-1, Native Country-2, Race-1, Workclass-2>	3	97.69	96.77
		4	97.16	96.98
		5	96.73	96.22
		6	96.21	95.11

Table 6.8: Comparison of proposed approach with existing k-anonymization approach (Execution Time)

Dataset	Quasi-identifier Generalization	value of k	Multi-Iterative K-Anonymization based	K-Anonymization based
			Execution Time (min)	Execution Time (min)
Bank Marketing	<Age-2,Job-2,Marital-1>	3	5:08:29	9:07:28
		4	5:17:13	9:14:09
		5	5:23:29	9:03:21
		6	3:27:45	7:35:12
Adult	<Age-2,Occupation-2, Sex-1, Relationship-1, Native Country-2, Race-1, Workclass-2>	3	3:02:18	6:26:05
		4	3:03:20	6:17:31
		5	3:50:23	6:42:21
		6	5:45:55	10:03:33

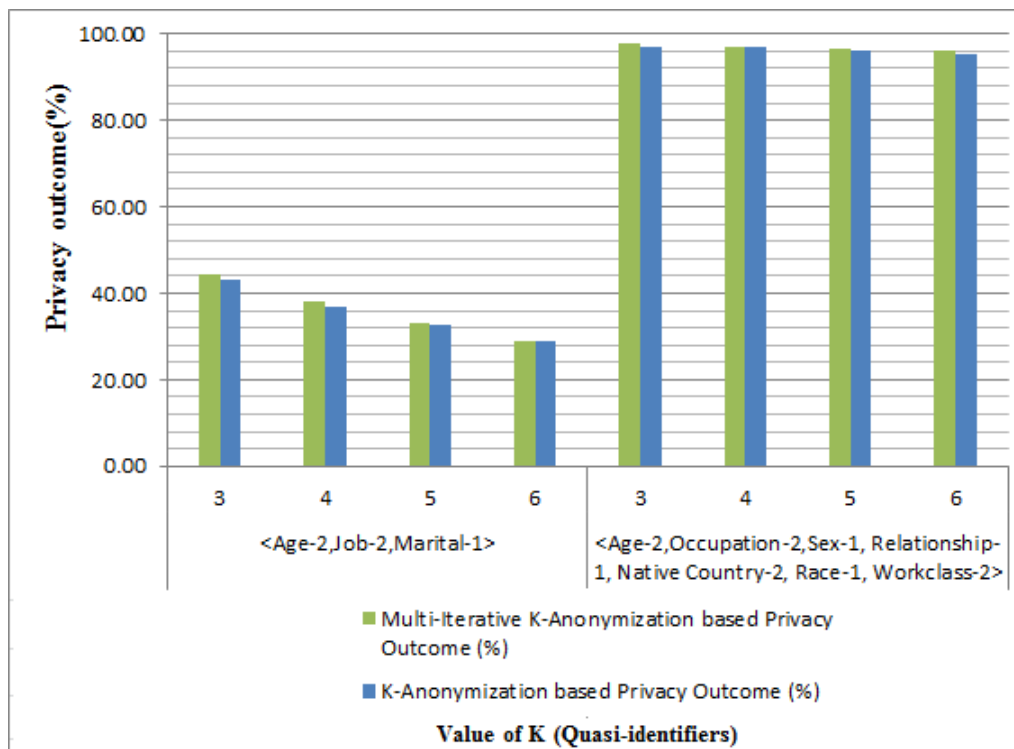


Figure 6.19: Comparison of proposed approach with k-Anonymization

6.5 Summary

Frequent exposure of private data while engaging in data-analysis activities has led to great threats to data privacy. Even though, such data are important assets to corporate decision-making; proprietors distributing information for data examination may prevent through data privacy concerns. The proposed research work tries to find out solutions for this growing concern. Several algorithms have been proposed to understand the characteristics of the data set and perturb either sensitive attribute values or keep sensitive attribute's values unchanged and anonymized quasi-identifier's values. Sensitive drift based data perturbation in stream mining approach minimizes the information loss like other proposed approaches of data perturbation. Under this research work we have also carried out a detailed analysis of data anonymization variants and the proposed heuristic-based hybrid privacy preserving k-anonymization with data perturbation. Privacy gain has been calculated before rescan of un-anonymized data set to get knowledge about subsequent level of best-fitting generalization that leads to minimum loss of information and better protection to an individual's privacy. Proposed method is more suitable for those data sets which contain multiple sensitive attributes.

Chapter 7

Conclusion, Limitations and Future Scope

In this section, we present the Conclusion, Limitations and Future Scope of the proposed approaches.

7.1 Conclusion

Protecting individual privacy throughout the method of data mining poses challenges to data mining community. Repeated disclosure of personal information, even while engaging in data analysis process, has led to great threats to data privacy. We know that data is a key property to corporate decision making through analyzing it and on the other side, dataset holders want to protect sensitive data by applying privacy on data during the analysis. The emergence of the new class of high-speed data-intensive streaming applications have opened the door for research in the data stream mining arena but data stream mining algorithms are not vigilant enough on data privacy.

Proposed research work makes an effort to find out solutions for this growing concern. A good number of algorithms have been suggested that understand the characteristics of the data set and perturb either sensitive attribute values or keep sensitive attribute values unchanged and anonymized quasi-identifiers values. Proposed Sensitive Drift based Perturbation approach in data mining works on tuple value and perturbs the sensitive attributes using the sliding window concept. Proposed approach shows improved results among other existing methods.

In this approach, the focus is on overall data characteristics and based on that each instance has been assigned Tuple Value and use of sensitive drift value, which then can be used to perturbed data. Sliding window based approach for data stream perturbation is the easiest approach for all proposed approaches but, it is fast among all. Even in the case of outliers, data stream perturbation using a sliding window will obliterate its effect outside window size.

In present research work, we have carried out detailed analysis of data anonymization variants and proposed Heuristic based hybrid privacy preserving data stream mining approach using Perturbation and K-anonymization. In this proposed approach, perturbation has been applied based on Sensitive Drift and anonymization and it has been further applied based on Privacy earn computed in multi-iteration. The proposed algorithm has an advantage of re-scanning of the only subset of the original data set which is not k-anonymized. We have tested the proposed PRIVACYearn based algorithm against standard data set and found that respondent's privacy has been protected with a small amount of data loss which occurs due to anonymization. Classification results with original data sets versus anonymized data sets over Naive Bayes algorithm have been compared. This algorithm also protects the sensitive data by converting the original sensitive value to perturb data value. proposed approach shows favorable results among other proposed approaches of data perturbation and anonymization.

Proposed hybrid Geometric based data perturbation on stream data perturbs the data using rotation, scaling and translation. We also find out the best suitable transformation order which will give the highest information gain among all other transformation orders. In our proposed approach, the accuracy of privacy depends on the security angle, a sequence of translation, scaling, and rotation. Results of proposed algorithm illustrate desirable privacy level which has been accomplished through equitable accuracy in closely all cases which are tested. Perturbed data stream has been measured through the ratio of occurrences of the dataset which are misclassified with the outcome of real dataset clustering. The proposed approach also provide favorable results.

7.2 Limitations

In this work, we considered the tradeoff between privacy and utility when mining streaming data and examined algorithms that allow the better tradeoffs. The proposed work is suitable for the numeric dataset and further improvement is required to make it work for nominal dataset also. Proposed work aims to provide privacy to individual's while publishing data to the outside world. This method is accepted for privacy-preserving data publishing unlike other proposed approaches based on data perturbation but this method is less suitable for data mining. Algorithms presented with this research have been tested against standard datasets with different parameters settings before concluding the accuracy of methods. The accuracy may vary if the dataset with outliers and/or significant missing values are used as input.

7.3 Future Scope

Proposed work has used several statistic based methods which are appropriate to numerical sensitive attribute values only. Still protecting the privacy of nominal attribute values is a challenging task. Protecting multiple sensitive attributes of a single dataset with minimum information loss and limited dataset scanning is one more open issue that needs further attention. Further work can be extended to propose a common framework to provide privacy on sensitive values and quasi-identifiers via learning characteristics of data sets and proposing suitable approaches for data perturbation and anonymization.

Appendix A

Test cases with standard data sets applied over proposed algorithms

The rationale behind selecting different blend is to compare accuracy of developed algorithms using data sets which are selected from sources widely followed by researchers. Datasets of binary class values as well as multiple class values have been used. More than one sensitive attributes from same data sets have been perturbed using developed algorithms and resultant average has been considered to evaluate the performance. Sufficient number of transactions have been used by changing input parameters value like number of clusters, set of quasi-identifiers, sliding window size to test algorithms.

[1]	Dataset:	COVERTYPE ((S. Moro and Rita))
	Source:	UCI machine learning repository
	No. of Instances:	5,81,012
	Class values:	{0,1,2,3,4,5,6}
	Test cases:	65,000 instances selected randomly, K-Mean clustering algorithm with different K and different sliding window size (w)
	Sensitive Attributes	Elevation, Aspect, Slope
[2]	Dataset:	ADULT (Kohavi and Becker)
	Source:	UCI machine learning repository
	No. of Instances:	48,842

	Class values:	{>50K, <=50K}
	Test cases:	48842 instances selected randomly, K-Mean clustering algorithm with different K and different sliding window size (w)
	Sensitive Attributes	Income
[3]	Dataset:	BANK MARKETING (Jock A. Blackard)
	Source:	UCI machine learning repository [110]
	No. of Instances:	45,211
	Class values:	{Yes, No}
	Test cases:	45,211 instances, K-Mean clustering algorithm with K=2, 3, 4, 5 and sliding window size w = 2000, 3000
	Test cases:	45,211 instances, Naive Bayes classification algorithm with anonymization k=2, 3, 4 and available quasi-identifiers q = 2, 3
	Sensitive Attributes	Age, Balance
[4]	Dataset:	ELECTRIC NORM (M. Harries)
	Source:	MOA Data sets
	No. of Instances:	45,312
	Class values:	{Up, Down}
	Test cases:	45,312 instances, K-Mean clustering algorithm with different K and different sliding window size (w)
	Sensitive Attributes	Nswprice, Nswdemand
[5]	Dataset:	AIRLINES (Ikonomovska)
	Source:	MOA Data sets
	No. of Instances:	5,39,383
	Class values:	{0,1}
	Test cases:	80,000 instances, K-Mean clustering algorithm with different K and different sliding window size (w)
	Sensitive Attributes	Flight
[6]	Dataset:	AGRAWAL (Agrawal, Imielinski, and Swami)
	Source:	Synthetic Dataset WEKA

	No. of Instances:	50,000
	Class values:	{0,1}
	Test cases:	50,000 instances, K-Mean clustering algorithm with different K and different sliding window size (w)
	Sensitive Attributes	Salary, Age

Index

- Certificate, i
- Conclusion, 123
- Cryptographic Based Methods, 34
- Data Mining, 11
- Data Mining - Privacy issues, 4
- Data Perturbation, 19
- Data Stream Classification, 14
- Data Stream Clustering, 16
- Data Stream Generators, 22
- Data Stream Mining, 12
- Data Stream Model, 12
- Declaration, ii
- Future Scope, 125
- GDP based PPDSM Algorithm, 47
- Heuristics Based Methods, 31
- Hybrid PPDSM Algorithm, 95
- K-Anonymity, 20
- Limitations, 125
- Massive Online Analysis, 21
- MOA features, 22
- Personalized privacy preservation, 33
- Precision, 23
- Privacy preservation based on utility, 33
- Privacy Preserving in Data Mining, 3
- Privacy Preserving in Data Stream Mining, 6
- Quasi-Identifiers, 19
- Recall, 23
- Research Issues & Challenges, 44
- SD-Perturbation, 73
- Summary of Major Research Contributions in PPDM and PPDSM, 37

Bibliography

- Aggarwal, Charu C. *Data streams: models and algorithms*. Vol. 31. Springer Science & Business Media, 2007.
- Aggarwal, Charu C and S Yu Philip. "A condensation approach to privacy preserving data mining." *International Conference on Extending Database Technology*. Springer, 2004. 183–199.
- Aggarwal, Charu C and Philip S Yu. "On variable constraints in privacy preserving data mining." *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 2005. 115–125.
- Aggarwal, Charu C, et al. "-A Framework for Clustering Evolving Data Streams." *Proceedings 2003 VLDB Conference*. Elsevier, 2003. 81–92.
- Agrawal, Dakshi and Charu C Aggarwal. "On the design and quantification of privacy preserving data mining algorithms." *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2001. 247–255.
- Agrawal, R., T. Imielinski, and A. Swami. "Database Mining: A Performance Perspective." *IEEE Transactions on Knowledge and Data Engineering* 5.6 (1993). Special issue on Learning and Discovery in Knowledge-Based Databases: 914–925. <<http://www.almaden.ibm.com/software/quest/Publications/ByDate.html>>.
- Agrawal, Rakesh, Alexandre Evfimievski, and Ramakrishnan Srikant. "Information sharing across private databases." *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. ACM, 2003. 86–97.
- Agrawal, Rakesh and Ramakrishnan Srikant. "Privacy-preserving data mining." *ACM Sigmod Record*. ACM, 2000. 439–450.

- Agrawal, Shipra and Jayant R Haritsa. "A framework for high-accuracy privacy-preserving mining." *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, 2005. 193–204.
- Babcock, Brian, et al. "Models and issues in data stream systems." *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2002. 1–16.
- Bayardo, Roberto J and R Agrawal. "Data privacy through optimal k-anonymization." *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, 2005. 217–228.
- Bifet, Albert and Richard Kirkby. "Data stream mining a practical approach." (2009).
- Bifet, Albert, et al. "Moa: Massive online analysis." *Journal of Machine Learning Research* 11.May (2010): 1601–1604.
- Blum, Avrim, et al. "Practical privacy: the SuLQ framework." *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005. 128–138.
- Bradley, Paul S, Usama M Fayyad, Cory Reina, et al. "Scaling Clustering Algorithms to Large Databases." *KDD*. 1998. 9–15.
- Cao, Feng, et al. "Density-based clustering over an evolving data stream with noise." *Proceedings of the 2006 SIAM international conference on data mining*. SIAM, 2006. 328–339.
- Cao, Jianneng, et al. "CASTLE: A delay - constrained scheme for k s-anonymizing data streams." *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008. 1376–1378.
- Chang, Joong Hyuk and Won Suk Lee. "Finding recent frequent itemsets adaptively over online data streams." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003. 487–492.
- Chao, Ching-Ming, Po-Zung Chen, and Chu-Hao Sun. "Privacy-preserving classification of data streams." *£* 12.3 (2009): 321–330.
- Chao, Po-Zung, Ching-Ming Sun, and Chu-Hao Chen. "Privacy Preserving Clustering of Data Streams." *£* 13.3 (2010): 349–358.
- Chaudhry, Nauman, Kevin Shaw, and Mahdi Abdelguerfi. *Stream data management*. Vol. 30. Springer Science & Business Media, 2006.

- Chen, Keke and Ling Liu. "A random rotation perturbation approach to privacy preserving data classification." (2005).
- . "Privacy-preserving multiparty collaborative mining with geometric data perturbation." *IEEE Transactions on Parallel and Distributed Systems* 20.12 (2009): 1764–1776.
- Chen, Keke, Gordon Sun, and Ling Liu. "Towards attack-resilient geometric data perturbation." *proceedings of the 2007 SIAM international conference on Data mining*. SIAM, 2007. 78–89.
- Chhinkaniwala, Hitesh and Sanjay Garg. "Tuple value based multiplicative data perturbation approach to preserve privacy in data stream mining." *arXiv preprint arXiv:1306.1334* (2013).
- Chu, Fang. "Mining techniques for data streams and sequences." Diss. University of California, Los Angeles, 2005.
- David J Hand, H. Mannila, P. Smyth. "Principles of Data Mining." 7 (2001).
- Domingo-Ferrer, Josep and Josep Maria Mateo-Sanz. "Practical data-oriented microaggregation for statistical disclosure control." *IEEE Transactions on Knowledge and data Engineering* 14.1 (2002): 189–201.
- Domingos, Pedro and Geoff Hulten. "Mining high-speed data streams." *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000. 71–80.
- Du, Wenliang and Zhijun Zhan. "Building decision tree classifier on private data." *Proceedings of the IEEE international conference on Privacy, security and data mining- Volume 14*. Australian Computer Society, Inc., 2002. 1–8.
- Dutta, Haimonti, et al. "Analysis of privacy preserving random perturbation techniques: further explorations." *Proceedings of the 2003 ACM workshop on Privacy in the electronic society*. ACM, 2003. 31–38.
- Ester, Martin, et al. "Incremental clustering for mining in a data warehousing environment." *VLDB*. Citeseer, 1998. 323–333.
- Evfimievski, Alexandre, et al. "Privacy preserving mining of association rules." *Information Systems* 29.4 (2004): 343–364.
- Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.

- Feelders, Ad, H Daniels, and Marcel Holsheimer. "Methodological and practical aspects of data mining." *Information & Management* 37.5 (2000): 271–281.
- Feigenbaum, Joan, et al. "Secure multiparty computation of approximations." *ACM transactions on Algorithms (TALG)* 2.3 (2006): 435–472.
- Fienberg, S. E. "Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation." *Statistical Science* 21 (2006): 143–154.
- Fienberg, Stephen E and Julie McIntyre. "Data swapping: Variations on a theme by Dalenius and Reiss." *Journal of Official Statistics* 21.2 (2005): 309.
- Fisher, Douglas H. "Knowledge acquisition via incremental conceptual clustering." *Machine learning* 2.2 (1987): 139–172.
- Fung, Benjamin CM, Ke Wang, and S Yu Philip. "Anonymizing classification data for privacy preservation." *IEEE transactions on knowledge and data engineering* 19.5 (2007): 711–725.
- Fung, Benjamin CM, Ke Wang, and Philip S Yu. "Top-down specialization for information and privacy preservation." *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*. IEEE, 2005. 205–216.
- Gama, João, Ricardo Rocha, and Pedro Medas. "Accurate decision trees for mining high-speed data streams." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003. 523–528.
- Goldreich, Oded. *Foundations of cryptography: volume 2, basic applications*. Cambridge university press, 2009.
- Guo, Ling and Xintao Wu. "Privacy Preserving Categorical Data Analysis with Unknown Distortion Parameters." *Trans. Data Privacy* 2.3 (2009): 185–205.
- Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
- Hand, David J, Heikki Mannila, and Padhraic Smyth. *Principles of data mining (adaptive computation and machine learning)*. MIT press Cambridge, MA, 2001.
- Hulten, Geoff, Laurie Spencer, and Pedro Domingos. "Mining time-changing data streams." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001. 97–106.

- Ikonomovska, Elena. "MOA Datasets." <https://moa.cms.waikato.ac.nz/datasets/>, 2009.
- Jin, Ruoming and Gagan Agrawal. "Efficient decision tree construction on streaming data." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003. 571–576.
- Jock A. Blackard, Denis J. Dean, Charles W. Anderson. "UCI Machine Learning Repository." A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, <https://archive.ics.uci.edu/ml/datasets/bank+marketing>, 2012.
- Kadampur, Mohammad Ali, et al. "A noise addition scheme in decision tree for privacy preserving data mining." *arXiv preprint arXiv:1001.3504* (2010).
- Kamakshi, P and A Vinaya Babu. "Automatic detection of sensitive attribute in PPDM." *Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on*. IEEE, 2012. 1–5.
- Kamakshi, A Vinaya, P and Babu. "Preserving the privacy and sharing the data using classification on perturbed data." *International Journal on Computer Science and Engineering* 2.3 (2010): 860–864.
- Kantarciloglu, Murat and Chris Clifton. "Privacy-preserving distributed mining of association rules on horizontally partitioned data." *IEEE transactions on knowledge and data engineering* 16.9 (2004): 1026–1037.
- Kantarciloglu, Murat, Jaideep Vaidya, and C Clifton. "Privacy preserving naive bayes classifier for horizontally partitioned data." *IEEE ICDM workshop on privacy preserving data mining*. 2003. 3–9.
- Kargupta, Hillol, et al. "On the privacy preserving properties of random data perturbation techniques." *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003. 99–106.
- Karmakar, Mithun and Dhruba K Bhattacharyya. "Privacy Preserving Data Mining Using Matrix Algebraic Approach." *Journal of Convergence Information Technology* 4.3 (2009): 38–44.
- Kifer, Daniel and Johannes Gehrke. "Injecting utility into anonymized datasets." *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006. 217–228.

- Kohavi, Ronny and Barry Becker. "UCI Machine Learning Repository." Data Mining and Visualization Silicon Graphics, <http://mlr.cs.umass.edu/ml/datasets/Adult>, 1996.
- Kranen, Philipp, et al. "Self-adaptive anytime stream clustering." *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009. 249–258.
- LeFevre, Kristen, David J DeWitt, and Raghu Ramakrishnan. "Mondrian multidimensional k-anonymity." *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006. 25–25.
- Li, Jianzhong, Beng Chin Ooi, and Weiping Wang. "Anonymizing streaming data for privacy protection." *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008. 1367–1369.
- Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007. 106–115.
- Lindell, Yehuda and Benny Pinkas. "Privacy preserving data mining." *Journal of cryptography* 15.3 (2002).
- Liu, Kun, Hillol Kargupta, and Jessica Ryan. "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining." *IEEE Transactions on knowledge and Data Engineering* 18.1 (2006): 92–106.
- M. Harries, Gama. "MOA Datasets." Australian New South Wales Electricity Market, <https://moa.cms.waikato.ac.nz/datasets/>,
- Machanavajjhala, Ashwin, et al. "L-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.
- Mahdavi, Mehrdad and Hassan Abolhassani. "Harmony K-means algorithm for document clustering." *Data Mining and Knowledge Discovery* 18.3 (2009): 370–391.
- Malik, Majid Bashir, M Asger Ghazi, and Rashid Ali. "Privacy preserving data mining techniques: current scenario and future prospects." *Computer and Communication Technology (ICCCT), 2012 Third International Conference on*. IEEE, 2012. 26–32.
- Maloof, Marcus A and Ryszard S Michalski. "Incremental learning with partial instance memory." *Artificial intelligence* 154.1-2 (2004): 95–126.

- Maron, Oded and Andrew W Moore. "Hoeffding races: Accelerating model selection search for classification and function approximation." *Advances in neural information processing systems*. 1994. 59–66.
- Mishra, Nina and Mark Sandler. "Privacy via pseudorandom sketches." *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2006. 143–152.
- Muthukrishnan, Shanmugavelayutham, et al. "Data streams: Algorithms and applications." *Foundations and Trends® in Theoretical Computer Science* 1.2 (2005): 117–236.
- Nabar, Shubha U, et al. "Towards robustness in query auditing." *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 2006. 151–162.
- O'callaghan, Liadan, et al. "Streaming-data algorithms for high-quality clustering." *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE, 2002. 685–694.
- Oliveira, Stanley RM and Osmar R Zaiane. "Achieving privacy preservation when sharing data for clustering." *Workshop on Secure Data Management*. Springer, 2004. 67–82.
- Ordonez, Carlos. "Clustering binary data streams with K-means." *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, 2003. 12–19.
- Ponnavaikko, M and E Poovammal. "Task independent privacy preserving data mining on medical dataset." *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*. IEEE, 2009. 814–818.
- Poovammal, E and M Ponnavaikko. "An improved method for privacy preserving data mining." *Advance Computing Conference, 2009. IACC 2009. IEEE International*. IEEE, 2009. 1453–1458.
- Rizvi, Shariq J and Jayant R Haritsa. "Maintaining data privacy in association rule mining." *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 2002. 682–693.

- S. Moro, P. Cortez and P. Rita. "UCI Machine Learning Repository." Colorado State University, Remote Sensing and GIS Program, Department of Forest Sciences, <https://archive.ics.uci.edu/ml/datasets/covertypes>, 1998.
- Samarati, Pierangela. "Protecting respondents identities in microdata release." *IEEE transactions on Knowledge and Data Engineering* 13.6 (2001): 1010–1027.
- Sato, Masa-Aki and Shin Ishii. "On-line EM algorithm for the normalized Gaussian network." *Neural computation* 12.2 (2000): 407–432.
- Schlimmer, Jeffrey C and Douglas Fisher. "A case study of incremental concept induction." *AAAI*. 1986. 496–501.
- Seidman, Claude. *Data mining with Microsoft SQL Server 2000 technical reference*. Microsoft Press, 2001.
- Singh, Meena Dilip, P Radha Krishna, and Ashutosh Saxena. "A privacy preserving jaccard similarity function for mining encrypted data." *TENCON 2009-2009 IEEE Region 10 Conference*. IEEE, 2009. 1–4.
- Street, W Nick and YongSeog Kim. "A streaming ensemble algorithm (SEA) for large-scale classification." *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001. 377–382.
- Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557–570.
- Thuraisingham, B. "Data mining, national security, privacy and civil liberties." *SIGKDD Explorations Newsletter* 4(2) (2002): 1–5.
- Tu, Li and Yixin Chen. "Stream data clustering based on grid density and attraction." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3.3 (2009): 12.
- Utgoff, Paul E. "Incremental induction of decision trees." *Machine learning* 4.2 (1989): 161–186.
- Vaidya, Jaideep and C Clifton. "Privacy-preserving k-means clustering over vertically partitioned data." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003. 206–215.

- Vaidya, Jaideep and Chris Clifton. "Privacy preserving association rule mining in vertically partitioned data." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002. 639–644.
- Vaidya, Jaideep, et al. "Privacy-preserving decision trees over vertically partitioned data." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2.3 (2008): 14.
- Vaidya, Chris, Jaideep and Clifton. "Privacy preserving naive bayes classifier for vertically partitioned data." *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 2004. 522–526.
- Verykios, Vassilios S, et al. "State-of-the-art in privacy preserving data mining." *ACM Sigmod Record* 33.1 (2004): 50–57.
- Wong, Raymond Chi-Wing, et al. " (α, k) -anonymity: an enhanced k-anonymity model for privacy preserving data publishing." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006. 754–759.
- Wu, Chai Wah. "Privacy preserving data mining with unidirectional interaction." *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*. IEEE, 2005. 5521–5524.
- Xiao, Xiaokui and Yufei Tao. "Personalized privacy preservation." *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, 2006. 229–240.
- Xu, Jian, et al. "Utility-based anonymization using local recoding." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006. 785–790.
- Yao, Andrew Chi-Chih. "How to generate and exchange secrets." *Foundations of Computer Science, 1986., 27th Annual Symposium on*. IEEE, 1986. 162–167.
- Zhan, Justin. "Privacy-preserving collaborative data mining." *IEEE Computational Intelligence Magazine* 3.2 (2008).
- Zhang, Gaofeng, Yun Yang, and Jinjun Chen. "A historical probability based noise generation strategy for privacy protection in cloud computing." *Journal of Computer and System Sciences* 78.5 (2012): 1374–1381.

- Zhang, Junwei, et al. "Kids: k-anonymization data stream base on sliding window." *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*. IEEE, 2010. V2–311.
- Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." *ACM Sigmod Record*. ACM, 1996. 103–114.
- Zhang, Xiaolin and Hongjing Bi. "Research on privacy preserving classification data mining based on random perturbation." *Information Networking and Automation (ICINA), 2010 International Conference on*. IEEE, 2010. V1–173.
- Zhang, XuePing, YanXia Zhu, and Nan Hua. "Privacy parallel algorithm for mining association rules and its application in HRM." *Computational Intelligence and Design, 2009. ISCID'09. Second International Symposium on*. IEEE, 2009. 296–299.
- Zhong, Sheng, Zhiqiang Yang, and Rebecca N Wright. "Privacy - enhancing k - anonymization of customer data." *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005. 139–147.
- Zhou, Bin, et al. "Continuous privacy preserving publishing of data streams." *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*. ACM, 2009. 648–659.