ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Performance Evaluation of Anonymized Data Stream Classifiers

[1] Aradhana Nyati, [2] Divya Bhatnagar

[1, 2] Department of Computer Science and Engineering,
Sir Padampat Singhania University,
Bhatewar, Udaipur, Rajasthan, India

Abstract - Data stream is a continuous and changing sequence of data that continuously arrive at a system to store or process. It is vital to find out useful information from large enormous amount of data streams generated from different applications viz. organization record, call center record, sensor data, network traffic, web searches etc. Privacy preserving data mining techniques allow generation of data for mining and preserve the private information of the individuals. In this paper, classification algorithms were applied on original data set as well as privacy preserved data set. Results were compared to evaluate the performance of various classification algorithms on the data streams that had been privacy preserved using anonymization techniques. The paper proposes an effective approach for classification of anonymized data streams. Intensive experiments were performed using appropriate data mining and anonymization tools. Experimental result shows that the proposed approach improves accuracy of classification and increases the utility, i.e. accuracy of classification while minimizing the mean absolute error. The proposed work presents the anonymization technique effective in terms of information loss and the classifiers efficient in terms of response time anddata usability.

Keywords - *Data Mining, Privacy Preservation, Data Stream, Privacy Preservation Data Mining, Anonymization, Classification, ARX-Tool.*

1. Introduction

Data mining has involved gradually attention in recent years because it is very important to be able to find out useful information from immense amounts of data. Subsequently, various data mining techniques have been developed. Typical mining include association mining, classification, and clustering. These techniques help to find interesting patterns, regularities, and anomalies in the data. Data mining is often applied to fields such drug development, business, finance, education, sports and stock market, Retail etc. Besides, the rapid advance in Internet and communications technology has led to the emergence of data streams. With today's information explosion, data not only are stored in large amount but also grow rapidly over time. Data stream has different characteristics of data collection to the traditional database model [2]. In data streams, data has timing preference, data distribution constantly keeps changing with time, data is enormous in amount, data flows in and out with fast speed and data requires immediate response[3]. Data streams are characterized as being unbounded, continuously arriving at a high speed rate and typically being scanned once [1]. The data mining techniques studied so far have been transformed from traditional static data mining to dynamic data stream mining due to the consecutive, rapid, temporal and unpredictable properties of data streams [2, 4,5]. Traditional algorithms are not appropriate for stream data due to large scale. These are designed for the static data-base. If the data changes, it would be a need for rescanning the database, which leads to more computation time and inability to promptly respond to the user [6]. These data sets need to be analyzed for finding patterns which help us in segregating anomalies and forecasting future behavior. So classification is required.The classification model is a representation of classification rules, decision trees, neural networks, or mathematical computation which is used for classification. The objective of data classification is to predict the (categorical) class labels of a given data tuples based on a training data set and to develop a model of classifier.

The power of data mining tools to extract hidden information from datasetsrequired increased data collection efforts by companies and government agencies. Naturally this raised privacy concerns about collected data [16]. In response to that, data mining researchers started to address privacy concerns. Special data mining techniques were developed under the framework of privacy preserving data mining. Opposed to regular data mining techniques, privacy preserving data mining can be applied

to databases without violating the privacy of individuals [7]. The aim of privacy-preserving data stream classification is to construct accurate classifiers without revealing private information.

2. Review of Literature

Fung et. al. proposed k-anonymization solution for classification. They conducted intensive experiments to evaluate the impact of anonymization on the classification on future data[1]. Su et. al. proposed a new associative classification algorithm for data streams. AC-DS based on the estimation mechanism of the Lossy Counting and landmark window model. This paper introduced a kind of mining association of data stream classification. If a concept drift takes place in it, this will not output an accurate result[8]. Ringneet.al. suggested the concept of moments to preserve the privacy of data streams along with compression of data. This technique was suitable for univariate data stream[9]. Chhinkaniwalaet.al.proposed an approach for privacy-preserving classification of data streams, which consists of two steps: data streams preprocessing and data streams mining. The classification result of perturbed data set using proposed algorithms shows data privacy with minimal information loss.

Disadvantage of proposed algorithms is that it can perturb sensitive attributes with numerical values only[10]. Pramod and Vyas discussed some of the issues of windowing concept for online stream mining to develop an effective, performance oriented algorithm. They discussed different approaches like Landmark Window Concept, sliding window and Damped Window Concept[11]. Trambadiya and Bhanodia proposed a Heuristic approach to preserve privacy with classification for stream data. They proposed the window approach algorithm to perturb the data and Hoeffding tree algorithm is applied on perturbed data. This approach preserves privacy and also improves processed to extract knowledge and build classification model for stream[12]. Patel et. al. proposed methods and algorithms which extends the existing process of data streams classification to achieve privacy preservation. He proposed following method for data perturbation: numeric values, non-numeric values, both numeric and non-numeric values.

There is scope for further improvement in proposed methods and algorithms for preserving privacy in data stream classification[13]. Dhivakar and Mohana discussed about anonymization,randomization, perturbation and distributed privacy preservation,the recent approaches involved in privacy preservation such as, each having its own advantages and disadvantages[14]. Patel and Shah focused on the existing techniques present in the field of privacy preserving data mining. It was found that all techniques perform in a different way as with respect to the type of data and the type of application or domain[15].It was observed that appropriate classifiers are require for non-stationary data. Privacy preservation complementation technique can be applied with cryptographic technique for more privacy gain and efficient classification techniques with anonymization can be further addressed for numeric and non-numeric values of attribute. Hence, efficient privacy preservation classifiers are required to achieve high accuracy, speed of mining, and increased level of privacy while maintaining the model-specific data utility.

3. Related Terms

3.1 Data Stream

A data stream is a real-time, continuous, and well-ordered sequence of items. It is not possible to control the order in which items arrive, nor achievable to locally store a stream in its entirety.

3.2 Classification

The objective of data classification is to predict the (categorical) class labels of a given data tuples based on a training data set and to develop a model of classifier.

Precision: which is defined as proportion of instances that are truly of a class divided by the total instances classified as that class.

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad (1)$$

where, tp= No. of examples predicted positive that are actually positive and fp = No. of examples predicted positive that are actually negative.

Recall: Recall is defined as proportion of instances classified as a given class divided by the actual total in thatclass. Recall means how complete the results are.

$$\text{Recall} = \frac{t_p}{t_p + f_n} \quad (2)$$

Where f_n is No. of examples predicted negative that are actually positive.

F-measures: It is a measure that combine recall and precision which is given as below:

$$\text{F} - \text{measure} = \frac{(2 * \text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (3)$$

Mean Absolute Error: The MAE measures the average magnitude of the errors in a set of forecasts. It measures accuracy for continuous variables.

$$\text{Mean Absolute Error(MAE)} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (4)$$

where, fi is prediction value, y_i is true value

Kappa statistic: It measures the agreement of prediction with the true class, formulated as given below:

$$k = \frac{p_o - p_e}{1 - p_e} \quad (5)$$

where, p_o is the relative observed agreement, p_e is the hypothetical probability of chance agreement.

Accuracy: It is a measure to determine the utility of the dataset [17].

$$\text{Accuracy} = \frac{\text{Correctly classified instances}}{\text{Total number of Instance}} \text{x } 100 \quad (6)$$

3.3 Techniques of Privacy Preservation

Privacy preservation means how an individual controls that who has access to his personal information. There are various techniques discussed in PPDM i.e. anonymization based PPDM, perturbation based PPDM, randomized response based PPDM, condensation approach based PPDM, cryptography based PPDM. Anonymization means identifying information is removed from original data in order to protect personal or private information. There are many ways to perform data anonymization basically this method uses k-anonymization approach. Among the two methods of anaonymization, the basic concept of suppression based algorithm is to mask some attributes by special value whereas generalization based k-anonymization are replaced with more general values based on value hierarchy Graph (VGH) [18].

The following section briefly describe anonymization algorithms.

K-anonymity: K-Anonymity is a method for providing privacy preservation by ensuring that data cannot be displayed to an individual. The main purpose is to protect individual privacy. If each row in the table cannot be distinguished from at least other k-1 rows by only looking a set of attributes, then this table is K-anonymized on these attributes.

L-diversity: The drawback of k-anonymization due to the background knowledge attack can be removed by diversifying the values of sensitive attribute within a block. An equivalence class is said to have l-diversity if there are at least l well-represented values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

Population Uniqueness: The term "uniqueness" is used to characterize the amount of unique records in a data set. These statistical methods estimate characteristics of the overall population with probability distributions and evaluate a set of population uniqueness.

Average Reidentification Risk: Average reidentification risk, is a combination of a re-identification risks combined with k-anonymity. It can be used to protect datasets from attacks.

4. Methodology

4.1 The Framework

A framework has been developed to evaluate the performance of various classification algorithms on the data stream that has been privacy preserved using various anonymization techniques. Firstly, a data stream is generated to get a data set D. Various classification algorithms are applied on original data set (D).Then the privacy preserving anonymization techniques are applied to the data set D to obtain a privacy preserved data set (D'). D' is then classified by appropriate classification algorithms. The results of classification of original data set (D) and that of privacy preserved data set (D') are compared. Fig. 1 shows the block diagram which describes the framework of the methodology.

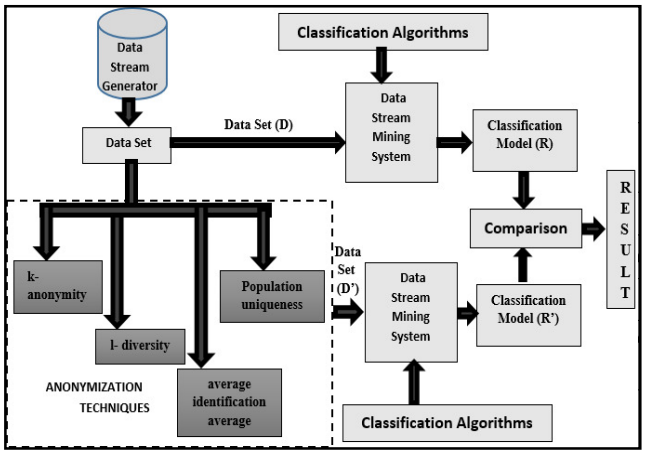

Fig. 1 Framework for classification of data stream using anonymization technique.

4.2. Experimental Setup

Data Set: Experiments have been carried out on the standard adult database from UCI [19] (University of California Irvine) machine learning repository with 30K instances and 15 attributes.9 attributes have been chosen for experimental purpose. The attributes were sex, age, race, marital-status, education, native-country, work class, occupation, and salary-class. The data set contained numerical and categorical attributes as well, which was suitable for generalization required. Salary was chosen as the class attribute. Sex, race, marital-status, education, native-country, work class, occupation were considered as quasi-identifier and age as sensitive attribute.

ARX-3.3.0: This tool was used for applying anonymization algorithms on the data stream [20].

WEKA(Waikato Environment for Knowledge Analysis) [21]: This tool was used for classification of data sets.

5. Results

5.1 Classification of Original Data Set

Classification algorithms were applied on original data set with WEKA. Outputs are tabulated in Table 1 and 2.

Table 1: Output of Classification of original data sets

Original Dataset	Class-ifier	MAE	Kappa statistics	Accuracy	Time (Sec)
Without privacy preservation	Jrip	0.27	0.499	82.28%	7.2
	J48	0.35	0.570	82.75%	0.3
	Naïve Bayes	0.22	0.513	81.63%	0.04

Table 2: output of classification of orignal data sets

Class-ifier	Precision Salary Class		Recall Salary Class		F-measure Salary Class	
	<=50	>50	<=50	>50	<=50	>50
Jrip	0.863	0.671	0.908	0.565	0.885	0.613
J48	0.862	0.692	0.917	0.557	0.889	0.617
Naïve Bayes	0.881	0.627	0.873	0.646	0.877	0.636

5.2 Anonymization of the Data Set

Out of l-diversity, k-anonymity, population uniqueness, and average reidentification risk applied on data set, minimum information loss was observed in population uniqueness.

5.3 Classification of Privacy Preserved Data Set

Classification algorithmswere applied on anonymized data set. Table 3 and 4and Fig.2, 3, 4, 5 and6 show the outputs.

Table 3: Result of classification on anonymized data set

Anonymi-zation Algori-thms	Classi-fier	MAE	Kappa Stats.	Accu-racy	Speed (Sec.)
l-diversity	Jrip	0.17	0.63	82.01%	5.18
	J48	0.16	0.65	82.16%	0.54.
	Naïve Bayes	0.15	0.64	81.83%	0.08
k-annony-mity	Jrip	0.16	0.63	82.38%	1.62
	J48	0.15	0.64	82.46%	0.01
	Naïve Bayes	0.14	0.61	81.69%	0.06
Popula-tion unique-ness	Jrip	0.16	0.63	83.34%	4.5
	J48	0.14	0.64	81.70%	0.01
	Naïve Bayes	0.14	0.60	81.69%	0.02
average-reidentific ation-risk	Jrip	0.17	0.60	82.01%	1.02
	J48	0.16	0.59	82.16%	0.02
	Naïve Bayes	0.15	0.57	79.64%	0.04

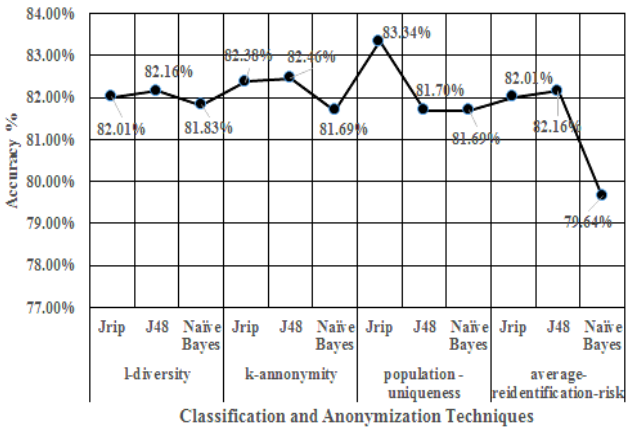

Fig. 2 Graphical representation of accuracy

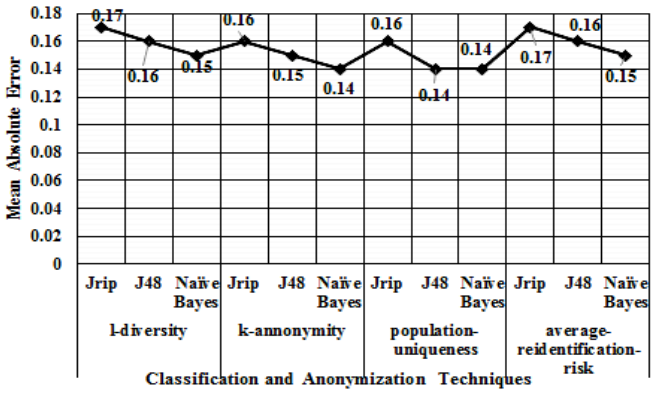

Fig. 3Graphical representation of mean absolute error obtained.

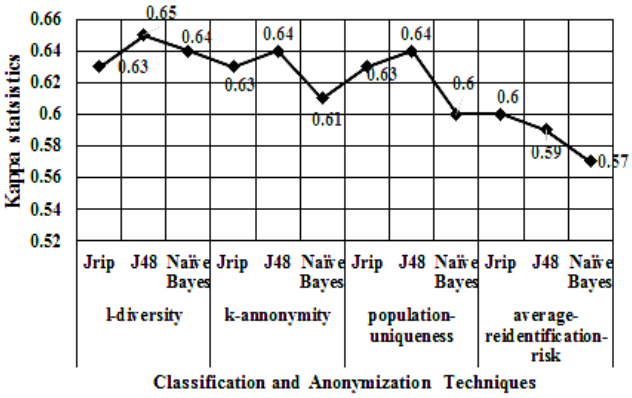

Fig. 4. Graphical representation of kappa statistics

Table 4: Output of parametric values of weka tool (privacy preserved data)

Anonym. Algorithms	Class-ier	Precision		Recall		F-measure	
		>50	<=50	>50	<=50	>50	<=50
l-diversity	Jrip	0.668	0.865	0.566	0.890	0.613	0.886
	J48	0.671	0.882	0.569	0.901	0.616	0.887
	Naïve Bayes	0.593	0.883	0.660	0.902	0.625	0.868
k-annonymity	Jrip	0.642	0.872	0.602	0.899	0.621	0.881
	J48	0.647	0.872	0.660	0.921	0.623	0.882
	Naïve Bayes	0.691	0.885	0.667	0.926	0.628	0.888
population uniqueness	Jrip	0.677	0.853	0.535	0.915	0.598	0.885
	J48	0.706	0.884	0.668	0.928	0.629	0.899
	Naïve Bayes	0.624	0.884	0.657	0.869	0.620	0.876
avg-iden.risk	Jrip	0.602	0.882	0.658	0.857	0.627	0.811
	J48	0.607	0.883	0.651	0.861	0.626	0.872
	Naïve Bayes	0.547	0.882	0.657	0.812	0.614	0.849

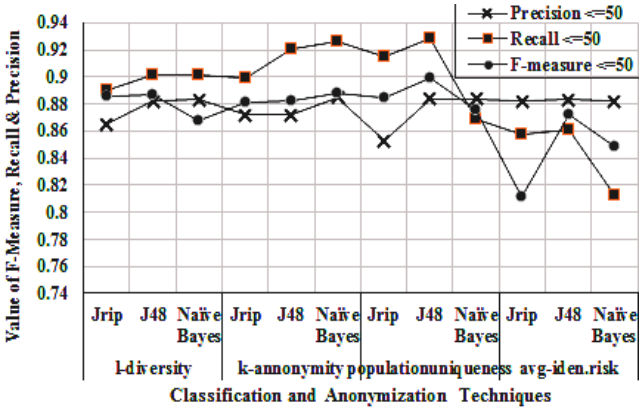

Fig.5. Graphical representation of parametric values(for the class salary<=50)

In order to compare the performance characteristics, the outputs were compared for two cases: One on privacy preserved dataset and the other on original dataset. Table 5 and 6 shows the results of comparison.

Table 5: Result of parametric values

Applying J-48	MAE	Kappa Stats	Accuracy	Speed (Sec)
Before Privacy Preservation	0.35	0.57	82.75%	0.03
After Privacy Preservation	0.15	0.64	82.46%	0.02

Table 6: Result of parametric values

Applying J-48		Precision	Recall	F-measure
Before Privacy Preservation	> 50K	0.692	0.557	0.617
	<=50K	0.862	0.917	0.889
After Privacy Preservation	> 50K	0.706	0.668	0.629
	<=50K	0.884	0.928	0.889

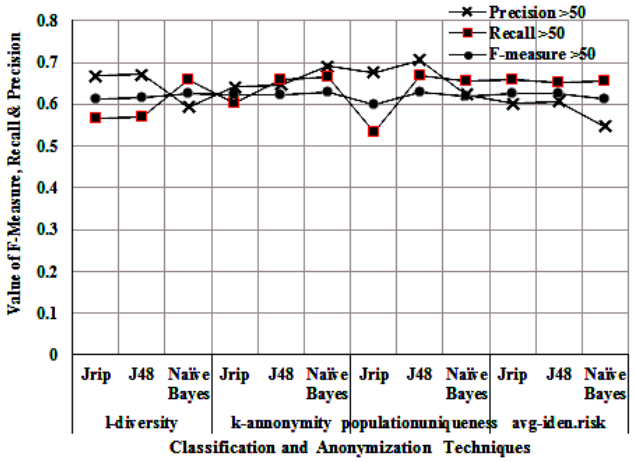

Fig. 6. Graphical representation of parametric values(for the class salary>50)

5. Conclusion and Future Scope

Applying the proposed framework, it was observed that amongst all anonymization techniques, minimum information loss was observed in population uniqueness. It was also observed that J-Rip when applied on the anonymization with k-anonymity and population uniqueness, J-48 applied on k-anonymity, Naive Bayes applied on k-Anonymity and population uniqueness produced good results. On the basis of these observations it is concluded that J-48 when applied on anonymization with population uniqueness and k-anonymity generated best results. Classification of anonymized data stream indicated increase in utility and decrease in mean absolute error. This work concludes with performance evaluation of data stream classifiers in which the data was privacy preserved using anonymization techniques. Yet there is enough scope in future to observe the performance of classifiers when other privacy preserving techniques such as randomization, perturbation etc., are applied for preserving the privacy of data streams.

Acknowledgments

We pay our special thanks for appreciative original work of all the authors of various technical papers which we have referred in initiation of the work without which it was very difficult to achieve successful completion. We also wish to put on record the word of appreciation for the developers of the tools, techniques and their easy access to learners and researchers.

References

[1] Benjamin C. M. Fung ,Wang K. , andPhilip S., "Anonymizing classification data for privacy preservation," IEEE Trans on Knowledge And Data Engineering, vol.19, 2007, pp.711-725.

[2] Aggarwal C. C. and Philip S.Yu. "A general survey of privacy-preserving data mining models and algorithms," Springer, vol.12, 2008, pp.11-52.

[3] Modi S. and Patel, "A. Privacy preserving data stream mining using two phase geometric data perturbation," International J for Scientific Research & Development, vol.3, 2015, pp.1115-1118.

[4] Chang J. H., and Lee W. S., "Finding recent frequent itemsets adaptively over online data stream," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., U.S.A., 2003, pp. 487-492.

[5] Cohen E. and Strauss M., "Maintaining time decaying stream aggregates," Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, San Diego, California, U.S.A., 2003, pp. 223-233.

[6] Vachhani N. and Vaghela B., "Geometric data transformation for privacy preserving on data stream using classification," IJIRCCE, vol.3, 2015, pp. 6013-6019.

[7] Patil S., Thakkar N., and Firoj S., "Secure Communication Using Privacy Preserving in a Data Mining," IJARCSSE, vol.5, 2015, pp.391-394.

[8] Li Su, Hong-yan Liu and Zhen-Hui Song, "A new classification algorithm for data stream," Intl J Modern Education and Computer Science, vol.4, 2011, pp. 32-39.

[9] Ringne A.G., Sood D. and Toshniwal D. , "Compression and privacy preservation of data streams using moments," Intl J of machine learning and computing, vol.1, 2011 pp.473-478.

[10] Chhinkaniwala H., Patel K. and Garg S., " Privacy preserving data stream classification using data perturbation techniques," Intl Conf on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012, pp. 1-8.

[11] Pramod S. and Vyas O. P. , "Data stream mining: a review on windowing approach," Global Journal of computer science and technology software and data engineering, vol.12, 2012, pp.27-30.

[12] Trambadiya T.J. and Bhanodia P., "A heuristic approach to preserve privacy in stream data with classification," Intl J of Engineering Research and Applications, vol.3, 2013, pp. 1096-1103.

[13] Patel M., Richariya P. and Shrivastava A., "Privacy preserving using randomization and encryption methods," Sch J Eng Tech, vol.1, 2013, pp.117-121.

[14] Dhivakar and Mohana, "A Survey on privacy preservation recent approaches and techniques," Intl J of Innovative Research in Computer and Communication Engineering, vol.2, 2014, pp.6559-6566.

[15] Brijlal P. and Shah, "An Overview of privacy preserving techniques and data accuracy," Intl J of Advance Research in Computer Science and Management Studies, vol.3, 2015 pp.135-140.

[16] Yin, Y. et. al , "Privacy Preserving Data Mining" in Data Mining, Springer, 2011, pp. 101-115.

[17] Devasena L., "Efficiency Comparison of Multilayer Perceptron and SMO Classifier for Credit Risk Prediction," Intl J of Advanced Research in Computer and Communication Engineering, 2014, vol.3, pp. 6155-6162.
[18] Patil M. and Ingale S., "Privacy Control Methods for Anonymous & Confidential Database Using Advance Encryption Standard," Intl J of Computer Science and Mobile Computing, 2013, vol.2, pp. 224-229.
[19] Lichman, M., "UCI Machine Learning Repository," 2013.
[20] http://arx.deidentifier.org/anonymization-tool.
[21] http://www.cs.waikato.ac.nz/ml/weka.

Aradhana Nyati, Research Scholar, Department of Computer Science and Engineering at Sir Padampat Singhania University, Udaipur, India has completed her post-graduation and graduation form MLS University Udaipur, Rajasthan in specialization with Information Technology. She has exposure to the academics and pursuing research in the field of data mining. Her area of research interest is data mining, network security, query processing and optimizing and dynamic data management.

Divya Bhatnagar is working as Professor in the department of Computer Science and Engineering in School of Engineering, Sir Padampat Singhania University, Udaipur, India. She holds 18 years of teaching and research experience. Her specialization areas include data mining and Neural Networks.