ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Data Mining on Privacy Protected Data using Naïve Bayes Algorithm

[1] Mr. Pankaj Jogi, [2] Mr.Shubham Purankar, [3]Ms. SamikshaPinge [4]Ms. Urvashi Ingale
[5]Ms. Neha Mahalle [6]Prof.V.S. Mahalle
[1]-[6] B.E. Final Year CSE

***Abstract* - Data mining has become an essential player in determining future business strategies. Data mining helps identifying patterns and trends from large amount of data, which can be used for reducing cost, increasing revenue and many more. With increased use of various data mining technologies and larger storage devices, amount of data collected and stored is significantly increased. This data contains personal information like credit card details, contact and residential information, etc. All these reasons have made it inevitable to concentrate on privacy of the data. In order to alleviate privacyconcerns, a number of techniques have recently been proposed to perform the data mining in privacy preserving way.**

***Keywords:*--- Naïve Bayes, data mining,data privacy, NB Algorithm**

I. INTRODUCTION

In recent years, the growing capacity of information storage devices has led to increased storing personal information about customers and individuals for various purposes. Data mining needs extensive amount of data to do analysis for finding out patterns and other information which could be helpful for business growth, tracking health data, improving services, etc. This information can be misused for many reasons like identity theft, fake credit/debit card transactions, etc. In order to alleviate these concerns a number of techniques have been proposed to perform data mining in privacy preserving way.

Some of the techniques are as following, Data Perturbation (hiding private data while mining patterns), Secure Multi-Party Computation (Building a model over distributed database without knowing others inputs), Knowledge Hiding (Hiding Sensitive rules), Privacy aware knowledge Sharing (do the data mining results themselves violet privacy). In data stream mining, the incoming data is continuous and real time and require algorithm that can change/alter the value before it may become available for data mining using classification or clustering techniques. It`s more difficult to provide privacy to real time data.

Data mining has both pros and cons. Issues of the data mining include a threat to privacy and security of the data. Sometimes data mining is done by a third-party service provider, which compromises the privacy of the data. Motivated by the privacy concerns of the data, new research area called privacy preserving data mining came into picture. Initial idea was to extend traditional data mining technologies by providing mask to sensitive information. But the key issues were how to modify the data and how to get the data stream mining result from the modified data. The goal is to achieve maximum accuracy for the intended data analysis task with less information loss, less response time and maximum privacy gain.

II. RELATED WORK

The one of the most important aspect of perturbation techniques is that the perturbed data generated by adding noise based on random data can be used to get an approximate inference very near to the original data. Hence the perturbed dataset can be accounted for mining accurate results. Perturbation techniques offer very low computation and communication cost.

Naive Bayes algorithm is one of the most effective methods in the field of text classification, but only in the large training sample set can it get a more accurate result. The requirement of a large number of samples not only brings heavy work for previous manual classification, but also puts forward a higher request for storage and computing resources during the computer post-processing. This paper mainly studies Naïve Bayes classification algorithm based on Poisson distribution model, and the experimental results show that this method keeps high classification accuracy even in small sample set.

III. METHODS AND MATERIALS

A. Need of Privacy Preservation

Privacy preserving while data mining has been major concern for long time now, with the improvements

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

in technologies, storage devices and advance software, now it is common to have large traditional databases and real time data as well. Few of the main sources of real time data are stock market, weather information coming from satellites, online transactions, internet traffic, telecommunication, etc.

The main difference between traditional databases and real time data is, data in statistical database doesn`t change with time and it can be stored and accessed later. Real time data is a stream of data which needs to be processed within that particular time frame and it is not stored as it is in huge amount. Both category datasets have different approach when it comes to mining. The main aspect in data mining applications dealing with sensitive information like educational, health care, financial, personal attributes, security, etc. is that it should preserve the privacy of the data. Specially, applications dealing with government records keeping, counter terrorism, self-defense, etc. For example, even though health organizations are allowed to release data after removing the identifiers like name, Identity numbers, address, etc. it is not considered safe enough since re-identification attacks have emerged which can link different public datasets to relocate the original subjects.[2] Sometimes organizations or private entities are not willing to distribute the sensitive information and also patterns detected by data mining systems can be used in a manner which violates the privacy of the individuals or organizations. These privacy constraints have led to exploring the new data mining techniques which protects the privacy of the data and also maintains the efficiency.

B. Geometric Data Perturbation

The idea behind using Geometric Data Perturbation algorithm is, because of its simplicity. Geometric perturbation is nothing but the enhancement to the rotation perturbation by coupling it with additional components like random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R £ X. It will be clear that by adding those additional components, Multiplicative perturbation for privacy preserving data mining geometric perturbation exhibits more robustness and provide efficiency in countering the attacks compared to normal rotation based perturbation.

For each attribute of G(X), let T be the translation, random rotation R, D be a Gaussian Noise and X be the original dataset. The value of the attribute G(X) can be found using following formula.

$$G(X) = R*X + T + D$$

Final G(X) will be as following.

ROTATED	Multiplication of ROTATION and ORIGINAL DATASET	Mean: SUM OF (original)/ no of elements	Translation Matrix: (original + mean)	PDF (gaussie n noise)	GX(geometirc data perturbation) (R.X+T+PD
5	5	3	4	0.76681	9.766813146
4	8		5		13.76681315
3	9		6		15.76681315
2	8		7		15.76681315
1	5		8		13.76681315

Datasets are required to evaluate data mining process. Adult Datasets are used for this project.

Attribute	Data type
Age	Numeric
Fnlwgt	Numeric
Work class	Text
Education	Text
Education num	Numeric
Marital Status	Text
Occupation	Text
Relationship	Text
Race	Text
Sex	Text
Capital gain	Numeric
Hours per week	Numeric
Native country	Text

(Table:1 Adult dataset)

C. Data Mining Algorithms and Techniques

Various algorithms and techniques like Artificial Intelligence, Classification, Nearest Neighbor Method, Clustering, Genetic Algorithm, Association Rules, Decision Trees, Regression, Neural Networks, etc., are used for performing data mining based on the desired outcomes, complexity and nature of the data.

a. Association Rule Learning: - This is also called market basket analysis or dependency modelling. It is used to discover relationship and association rules among variables.

b. Clustering: - This technique creates and discovers group of similar data items. This is also called unsupervised classification.

c. Classification: - This can classify data according to their classes i.e. put data in single group that belongs to a common class. This is also called supervised classification.

It is most commonly used data mining technique. Classification algorithms work on a basic principle of predicting certain outcome based on a type of input given. “A Classification algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations” .

d. Regression: - It tries to find a function that model the data with least errors.

e. Summarization: - It provides easy to understand and analysis facility through visualization, reports etc.

D. Knowledge Discovery Process

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

"Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD)". Modern computer systems are capable to save at an almost unimaginable rate and from wide range of data generating sources from point of sale machines to machines logging every single credit/debit card transactions, cash withdrawal and check clearances to, to space observing satellites. Here are some examples for better understanding of the volume of the data. "The current NASA earth observation satellite generates a terabyte of data every data".

Data mining is also known as Knowledge Discovery in Databases (KDD). While data mining and KDD are often treated as synonyms, data mining is actually a part of KDD process. Knowledge discovery in databases process consists of following steps:

Data cleaning: in this phase noise data and irrelevant data are removed.

Data Integration: in this phase multiple data sources may be combined in a common source.

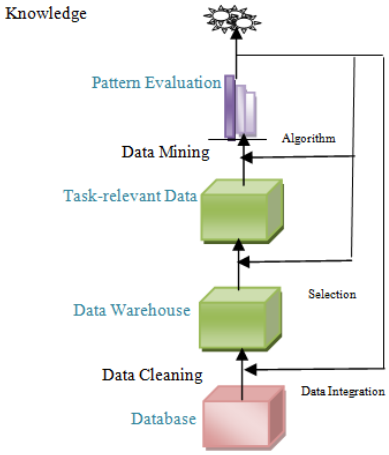
Data selection: this step decides the relevant data for analysis is decided and retrieved from collection.

Data transformation: selected data is transformed into forms appropriate for mining procedures.

Data mining: it is a crucial step where mining techniques are applied to extract patterns.

Pattern evaluation: valuable patterns are identified based on provided measures.

Knowledge representation: in a final phase, discovered knowledge is visually presented to the user. It helps user interpret data mining results. The following figure shows data mining step in knowledge discovery process:[3]

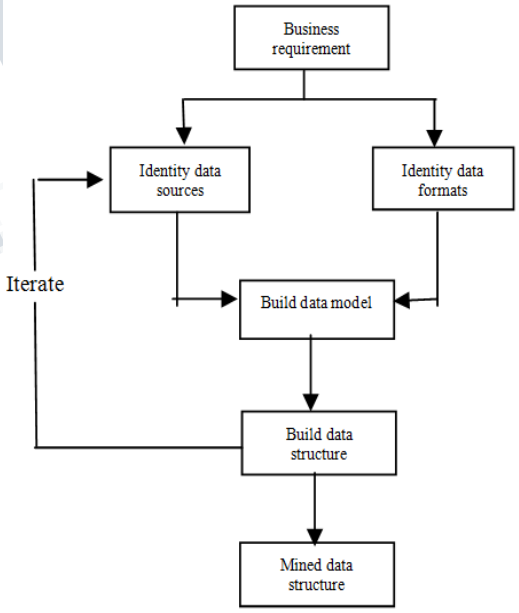


(Fig1: Data mining in the core of Knowledge Discovery process)

Large data has caused the use of more intense and complex techniques of data mining, partially because the size of data is too big and information varies in its content. With big data sets, merely getting relatively simple and straightforward statistics out of the system is not enough. For example, the country of 100 million populations, knowing that 10 million out of them live in a particular region is not enough. It`s crucial to know their age, economy background, gender, etc. to carry out various government schemes. Similarly, such detailed analysis can come handy for businesses to target potential customers.

Such business-driven needs have changed the simple data retrieval practices into more complex and detailed data retrieval techniques. The process of analysis and knowledge discovery is often iterative as it involves identification of different information that can be extracted. Data mining also requires the understanding of relating, mapping and clustering the one data to the other data to produce the result.

The figure shown below outlines the process of mining and model building for the business-driven data mining.



(Fig:2 Business-driven Data Mining)

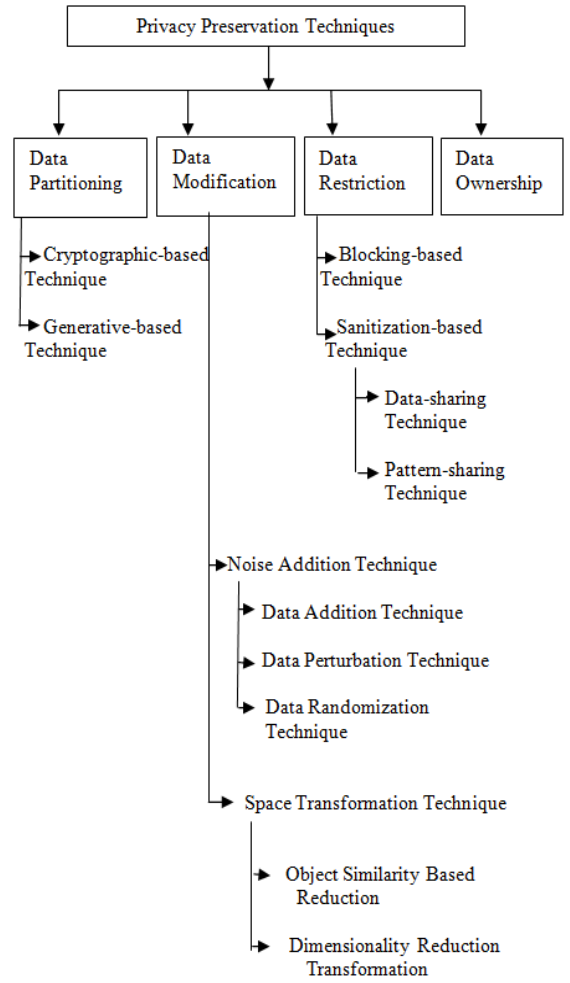
E. Privacy Preserving Data Mining Techniques

There are two dimensions of Privacy preserving in data mining [4]. i. Individual Privacy Preservation: This dimension mainly focuses on the privacy of the individuals or private entities.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

ii. Collective Privacy Preservation: As the name suggests, main area of this dimension is on the privacy of the overall organizations. Various techniques have been proposed varying from entire dataset modification to selective dataset modification. Most of privacy preserving data mining techniques can be classified in the following categories: Data Partitioning, data Modification, Data Restriction, Data Ownership. Following figure lists out few of privacy preserving data mining techniques practiced by various organizations dealing with the sensitive data [5].



(Fig.3 Privacy preserving data mining techniques)

The focus of this project is on Noise addition techniques, more on data perturbation techniques[4].

F. Naïve Bayes Algorithm

The Naïve Bayes is one of the classification type of algorithm used in data mining. It is based on conditional probability which uses Bayes theorem formula. The formula is used in order to calculate the probability by computing the frequency of values and the combinations from the previous data records or from databases [1].

The Bayes theorem computes the probability of occurring of event, given the probability of event occurred already. The formula for Bayes theorem is given as:

$$P(c \mid x) = \frac{P(x \mid c) * P(c)}{P(c)}$$

Where,

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

IV. EXPERIMENTAL STUDY

For the experiment purpose"Adult" datasets are used. As discussed earlier this algorithm has some limitations like, it can be applied only on numeric dataset and only on one attribute at a time.

Below shown tables demonstrate the results acquired by applying the Geometric Data Perturbation algorithm on two numeric attributes of each data set.

As mentioned earlier in the implementation steps, perturbed dataset D' is obtained by adding Gaussian Noise in the original dataset D and different factors of both the datasets are compared after applying two different algorithms on the original as well as perturbed datasets.

Here, WEKA tool is used to apply NB (Naïve Bayesian classifier algorithm) and. Below tables show the comparison between original values and modified values and their efficiencies.

	Age		Education	
	Original	Perturbed	Original	Perturbed
Correctly classified instances	0.8278	0.8246	0.8278	0.8194
Incorrectly classified instances	0.1721	0.1753	0.1721	0.1807
Time taken	0.02	0.02	0.02	0.03
Kappa Statistics	0.4992	0.4867	0.4992	0.485

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 3, March 2017

Mean Absolute error	0.182	0.1842	0.182	0.177
Root Mean Squared error	0.374	0.3786	0.374	0.3632
Relative absolute error	0.4868	0.4928	0.4868	0.4947
Root relative squared error	0.8652	0.8758	0.8652	0.8782

(Table:2 Adult Dataset)

From the table 2 we can see that after Applying NB (Naïve Bayesian) algorithm on both original and perturbed data, accuracy is nearly same in both scenarios.

This means with NB (Naïve Bayesian) algorithm on original adult data the accuracy of

Correctly classified instances	82.78%
Incorrectly classified instances	17.21%

This means with NB (Naïve Bayesian) algorithm on perturbed adult data the accuracy of

Correctly classified instances	82.46%
Incorrectly classified instances	17.53%

For this dataset, it proves that using Geometric data perturbation on dataset. The privacy of the original data can be preserved by little accuracy loss.
Results from both algorithms show that Geometric data perturbation can be a good alternative where accuracy is not utmost important for a data owner than the security of the data.

V. CONCLUSION AND FUTURE WORK

Geometric Transformation technique based on Multiplicative Data Perturbation approach has been applied for adding random noise to the original dataset to preserve privacy of sensitive attributes. This approach has been in direction to keep statistical relationship intact to mine useful results with perturbed data.It takes sensitive attributes as dependent attributes whereas, remaining attributes of dataset except class attribute are considered as independent attributes. Any calculations for adding tuple specific random noise is done only on dependent attributes of the dataset. The above framework uses Naïve Bayesian Classification algorithm to estimate the correct value of classification results from perturbed dataset over results from standard dataset. Accuracy of the results from the perturbed data will be less than the accuracy of the results from the original dataset. But, it is possible to achieve main objective of preserving the privacy of the sensitive info with less accuracy loss and the loss can further be minimized.

The paper uses Geometric Data Perturbation technique for numeric data only. The algorithm can be applied on non- numeric dataset using k-anonymization techniques. The algorithm can also be expanded to check real time data using stream analysis tool. The applied algorithm is working to extract single column value only and can be extended for more than one column at a time and also this algorithm can be applied to more classification as well as clustering algorithms. One of the future goals can also be to improve efficiency of the data mining of altered dataset and makeprivacy preserving more effective with minimal accuracy loss.

REFERENCES:

[1] Murat kantarcioglu and Jaideep Vaidya "PrivacyPreservation Naive Bayes Classifierfor Horizontally partitioned data". Purdue university.
[2] Dr. Lokanatha C. Reddy, "A review on Data mining from Past the Future", International Journal of Computer Applications (0975-8887) Volume 15- No. 7, February 2011.
[3] Usama Fayyad ,Gregory Piatetsfy-Shapiro, and Padhraic Smyth, "Data Mining to knowledge Discovery in Databases", AI Magazine Volume 17 Number 3 (1996).
[4] Kalyani M Raval,"Data Mining Technique", International Journal of Advance Research in Computer Science and Software Engineering Volume 2, Issue1, October 2012.
[5] Jharna Chopra, SampadaSatav "PRIVACY PRESERVATION TECHNIQUES IN DATA MINING", International Journal of Research in Engineering and Technology,ISSN: 2319-1163,Volume: 02 Issue: 04 | Apr-2013.