ANALYSIS OF VARIOUS TECHNIQUES AND KEY FIELDS OF PRIVACY PRESERVATION IN DATA MINING

S. INDUJA[1], N. REVATHI[2]

[1]Assitant Professor, Department of Computer Science, Tirupur Kumaran College for Women, Tirupur, India.

indujakmc@gmail.com

[2] Research Scholar, Department of Computer science, Tirupur Kumaran College for Women, Tirupur, India.

Revathividhya54@gmail.com

Abstract: Data mining is a useful data extraction from hidden knowledge of prediction information from a large data base. In recent years, data mining is considered a threat to privacy because electronic data maintained by companies is widespread. This fundamental data has led to increased concern about privacy. In recent years, many techniques have been proposed to modify or alter the data to protect privacy. The main purpose of this paper is to analyze the privacy-protecting technologies and key sectors that implement privacy-protective procedures.

Keywords: Privacy, Preserving, Data Mining, Clustering, Association, Application, Knowledge Discovery Database and Models.

I. INTRODUCTION

Data Mining is a database, with arrangements to extract hidden data from a large database. Sequencing lots of information packages and using complex calculations to select important data. Expect future methods and procedures for information mining equipment, allowing interactive companies to perform approved operating options. More information is collected by the amount of information being multiplied every year, and the Information Mining Motion has become indispensable to change this information into data [9]. Information technology was developed in a long analysis of experiments and improvements in the products. Data mining analytics works with data and the best techniques look at data greatly and use data collected as much as possible to obtain reliable results and results. Analysis process begins with a set of data, which uses a method to generate optimal representation of the structure of data acquired timely. Once acquired knowledge, a large data set can be extended to larger packages of data that can be assumed to have an assumption of pattern. Is again similar to a mining operation, where large amounts of low-sized goods are wiped out by the means of finding the value.

The following diagram summarizes certain stages / processes to identify data mining and knowledge innovation.

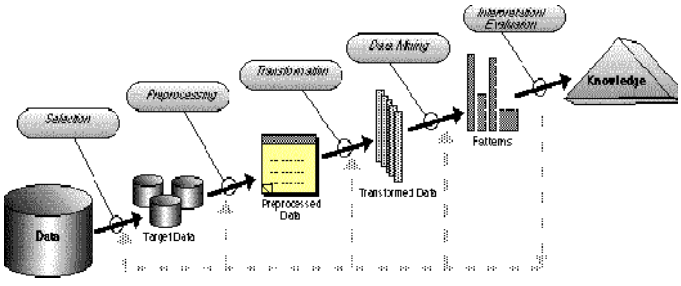

Figure 1: - KDD Process

The phases depicted start with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:

- **Selection** - Selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.

- **Preprocessing** - This is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying pregnancy. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0.
- **Transformation** - The data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.
- **Data mining -** This stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) *F*, a language *L*, and some measure of certainty *C* a pattern is a statement *S* in *L* that describes relationships among a subset *Fs* of *F* with a certainty *c* such that *S* is simpler in some sense than the enumeration of all the facts in *Fs*.
- **Interpretation and evaluation** - The patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

II. DATA MINING TECHNIQUES

Data mining is indirectly lacking, which is unknown and provides meaningful data from the database. It includes various technological approaches. Data mining creates a descriptive model or a forecast model. An explanatory model describes the general characteristics of the data in the database. A prediction model makes the current data reliable to predict. The predictive and descriptive model goal is to achieve a variety of data mining techniques as shown in figure 2

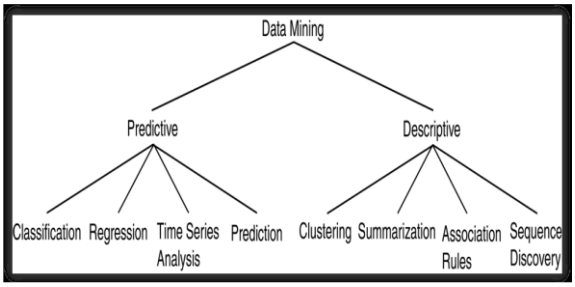

Figure 2 Data Mining Models

Classification: Classification based on assorted values (i.e., isolated, unordered). This technique based on supervised learning (that is, the output required for input is known) can not predict the following values. This is derived from the existing data and values (the class label). Conclusion The tree can be classified using the neural network. Mathematical formulas and classification rule (IF-Then).

Regression: Regression is used to map a data item to a real valued prediction variable. In other words, regression can be adapted for prediction. The target value is known in this technique.

Time Series Analysis: The statistical techniques are used in time series analysis and gives detail about data points which is dependent on time series. Time series forecasting is used to generate predictions of future events depend on past events.

Prediction: Prediction discovers the relationship between independent variables and dependent variables. It gives continuous value or ordered value (between some range).

Clustering: Clustering is a set of similar data object. The steering object is another clustering. It analyzes data objects without a subject label. For example, it involves a lack of supervision, a retailer can build various committees based on a customer base such as weekly purchase and regular purchase.

Summarization: Summarization is abstraction of data. It is set of relevant task. For example, long distance calls can be summarized total minutes, seconds and total cost of the call.

Association Rule: It is a data mining techniques which give a set of items and a huge collection of transaction in frequent item set. Association strives to discover patterns in data which are based upon

relationships between items in the same transaction. Association rule is used in the market based analysis to identify a set, or sets of products that consumers often purchase at the same time.

Sequence Discovery: It is used among data for uncovers relationship. It is set of object each associated with its own timeline of events.

III. PRIVACY PRESERVING DATA MINING

The proprietary data processing is a new research direction in data mining and statistical databases, and data mining can be analyzed to the side effects of data privacy. The key instrument of privacy is two times to protect the privacy. Firstly, important source data such as identifiers, names, addresses, etc. should be corrected or summed from the original database in order to obtain unnecessary data to compromise individual privacy. Secondly, the important knowledge that can be deducted from a database using data mining techniques should also be excluded because such knowledge can correctly compromise the secrets of the data, as we will point out.

The main purpose of the data mining agency is to create guidelines to alter the original data, which will be private after private data and private knowledge mining. Problem derived from unauthorized users from data released during confidential information is commonly called a "database assumption" problem. In this statement, we provide a classification and detailed description of the various technologies and methods created in the area of privacy and data mining.

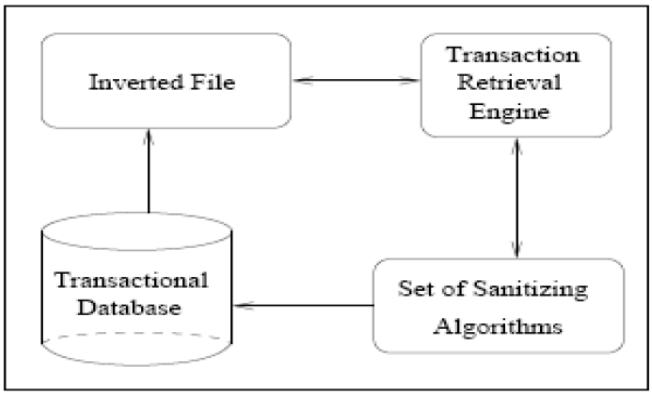

Figure 3: - Framework of PPDM

There are many approaches which have been adopted for privacy preserving data mining.

We can classify them based on the following dimensions:

- **Data distribution.**
- **Data modification.**
- **Data mining algorithm.**
- **Data or rule hiding.**
- **Privacy preservation.**

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization.

Methods of modification include:

- Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise),
- Blocking, which is the replacement of an existing attribute value with a “?”,
- Aggregation or merging which is the combination of several values into a coarser category,
- *S*wapping that refers to interchanging values of individual records, and
- Sampling, which refers to releasing data for only a sample of a population?

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design

of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms, into our future research agenda. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion" [3]. The last dimension which is the most important refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized.

The techniques that have been applied for this reason are:

- Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values.
- Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the Loss of functionality.

The PPDM algorithms can be first divided into two major categories [6], centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places. While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive.

The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding [4]. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data.

Currently, the PPDM algorithms are mainly used on the tasks of classification, association rule and clustering. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Classification is the process of finding a set of models (or functions) that describe

and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Clustering Analysis concerns the problem of decomposing or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.

PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution [2]. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

IV. KEY FIELDS OF PPDM

Privacy Preserving Data Mining (PPDM) [1] has risen to address this issue. The greater part of the methods for PPDM uses changed rendition of standard information mining calculations, where the adjustments typically utilizing great known cryptographic systems guarantee the obliged protection for the application for which the method was composed. Much of the time, the requirements for PPDM are saving precision of the information and the created models and the execution of the mining methodology while keeping up the protection demands. The few methodologies utilized by PPDM could be compressed as underneath:

- The information is changed before conveying it to the information excavator.
- The information is appropriated between two or more destinations, which participate utilizing a semi-fair convention to learn worldwide information mining results without uncovering any data about the information at their individual locales.
- While utilizing a model to order information, the arrangement results are just uncovered to the assigned party, who does not learn whatever else might be available other that the characterization results, yet can check for vicinity of specific principles without uncovering the guidelines.

The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and *k*-anonymity have been suggested in recent years in order to perform privacy-preserving data mining [3]. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar.

The key directions in the field of privacy-preserving data mining are as follows:

i) Privacy-Preserving Data Publishing:

These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization, *k*-anonymity, and *l*-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of

determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

ii) Changing the results of Data Mining Applications to preserve privacy:

In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

iii) Query Auditing:

Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries.

iv) Crypto-graphic Methods for Distributed Privacy:

In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

v) Theoretical Challenges in High Dimensionality:

Real data sets are usually extremely high dimensional and this makes the process of privacy-preservation extremely difficult both from a computational and effectiveness point of view. In, it has been shown that optimal *k*-anonymization is NP-hard [7]. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view [4], this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

V. CONCLUSION

This paper gives data mining that protects privacy and technologies involved in various fields. The idea of privacy to protect data mining is to extract information from active data. Applications use some data mining classification, suite, computation, and association rules and so on. So in future work we can review various classifications and clustering algorithm and its significance's and implemented with cryptography techniques..

VI. REFERENCES

[1] D.Aruna Kumari , Dr.K.Rajasekhar rao, M.suman " Privacy preserving distributed data mining using steganography "In Procc. Of CNSA-2010, **Springer Libyary**

[2] T.Anuradha, suman M,Aruna Kumari D "Data obscuration in privacy preserving data mining in Procc International conference on web sciences ICWS 2009.

[3] Agrawal, R. & Srikant, R.(2000). Privacy Preserving Data Mining. In Proc. of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, TX.

[4] Alexandre Evfimievski, Tyrone Grandison Privacy Preserving Data Mining. IBM Almaden Research Center 650 Harry Road, San Jose, California 95120, USA

[5] Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York, Springer, 2008.

[6] Oliveira S.R.M, Zaiane Osmar R., A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration, In Proceedings of the International Workshop on Privacy and Security Aspects of Data Mining in conjunction with ICDM 2004, Brighton, UK, November 2004.

[7] Flavius L. Gorgônio and José Alfredo F. Costa"Privacy-Preserving Clustering on Distributed Databases:A Review and Some Contributions

[8] D.Aruna Kumari, Dr.K.rajasekhar rao,M.Suman "Privacy preserving distributed data mining: a new approach for detecting network traffic using steganography" in international journal of systems and technology(IJST) june 2011.

[9] Binit kumar Sinha "Privacy preserving, and C. S. Yang, A Fast VQ Codebook Generation Algorithm via Pattern Reduction, *Pattern Recognition Letters*, vol. 30, pp. 653{660, 2009}

[10] C. W. Tsai, C. Y. Lee, M. C. Chiang Kurt Thearling, Information about data mining and analytic technologies http://www.thearling.com/

[11] K.Somasundaram, S.Vimala,"A Novel Codebook Initialization Technique for Generalized Lloyd Algorithm using Cluster Density", International Journal on Computer Science and Engineering, Vol. 2, No. 5, pp. 1807-1809, 2010.

[12] K.Somasundaram, S.Vimala, "Codebook Generation for Vector Quantization with Edge Features", CiiT International Journal of Digital Image Processing, Vol. 2, No.7, pp. 194-198, 2010.

[13] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino State-of-the-art in Privacy Preserving Data Mining in SIGMOD Record, Vol. 33, No. 1, March 2004.

[14] Quantization: A Review", IEEE Transactions on Communications, Vol. 36, No. 8, August 1988.

[15] Berger T, "Rate Distortion Theory", Englewood Cliffs, Prentice-Hall,NJ, 1971.

[16] A.Gersho and V.Cuperman, "Vector Quantization: APattern Matching Technique for Speech Coding", IEEE Communications, Mag., pp 15-21, 1983.

[17] "Privacy Preserving Data Mining - IBM Research: Almaden: San Jose

[18] D.Aruna Kumari, Dr.K.Rajasekhara rao, M.suman "Privacy Preserving Clustering in DDM using Cryptography"in TJ-RJCSE-IJ-06.