An Empirical Study on Privacy Preserving Data Mining

1. Md. Riyazuddin, Asst.Prof-IT, Muffakham jah College of Engineering and Technology,Hyderabad,India.
2.Dr.V.V.S.S.S.Balaram,Prof and HOD-IT, Sreenidhi Institute of Science and Technology,Hyderabad,India.
3. Md.Afroze,Asst.Prof-IT,Muffakham jah College of Engineering and Technology,Hyderabad,India.
4. Md.JaffarSadiq, Assoc.Prof-IT, Sreenidhi Institute of Science and Technology,Hyderabad,India.
5.M.D.Zuber,Asst.Prof-CSE,Hi-Point college of Engineering and Technology, Moinabad, Hyderabad, India.

Abstract: In modern years, advances in hardware expertise have lead to an increase in the competence to store and record personal data about consumers and individuals. This has lead to concerns that the personal data may be misused for a variety of purposes. In order to lighten these concerns, a number of techniques have newly been proposed in order to perform the data mining tasks in a privacy-preserving way. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes. So society have become increasingly indisposed to share their data, frequently resulting in individuals either refusing to share their data or providing incorrect data. Privacy preserving data mining has been studied extensively, because of the wide explosion of sensitive information on the global source. In this paper, we provide a review of methods for privacy and analyze the representative technique for privacy preserving data mining.

Keywords: Privacy preserving, Data Mining, Techniques, Analysis.

1. Introduction

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, 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. Privacy preserving data mining [9, 18], is a novel research direction in datasome way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the “database inference” problem. In this explosion, we provide a classification and an extended description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.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 algorithm to leverage this information. A number of techniques have been suggested in recent years in order to perform privacy preserving data mining [5, 9, 11, 15]. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domains marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine certain kinds of data without violating the data owners ’privacy .For example, how to mine patients ’private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns are increasing. Commercial concerns are also concerned with the privacy issue. Most organizations collect information about individuals for their own specific needs. Very frequently, however, different units within an organization themselves may find it necessary to share information.In such cases, each organization or unit must be sure that the privacy of the individual is not violated or that sensitive business information is not revealed .Consider, for example, a government, or more appropriately, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicating the necessity for further examination derives from a wide variety of sources such as police records; airports; banks; general government statistics; and passenger information records that generally include personal information (such as name and passport number); demographic data (such as age and gender); flight

information (such as departure, destination, and duration); and expenditure data (such as transfers, purchasing and bank transactions). In most countries, this information is regarded as private and to avoid intentionally or unintentionally exposing confidential information about an individual, it is against the law to make such information freely available.

The rest of the paper organized, which exploring the Taxonomy of Privacy Preserving Techniquesin section 2 , Assessment of Privacy Preserving Algorithms in section 3, section 4 explores the Distributed Privacy Preserving Data Mining, section 5 discuss the Evaluation of Privacy Preserving Algorithms and , section 6 revealed with the conclusion followed by references.

2. Taxonomy of Privacy Preserving Techniques

There are many methodologies which have been accepted for privacy preserving data mining. We can categorize them based on the following measurements:
•Data Distribution
•Data Modification
•Data Mining Algorithm
•Data or Rule hiding
•Privacy Preservation
The first dimension discusses 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 discusses 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 [16, 18]. 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,
•swapping 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 discusses 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 discusses 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".
The last dimension which is the most important, discusses 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.

2.1 Perturbation Methodology

The perturbation methodology works under the need that the data service is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. Since the method does not reconstruct the original data values but only distributions, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem such as classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. For example, Agrawal [3] develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton and Rizvi and Haritsa[4] develop methods for privacy-preserving association rule mining. While some clever approaches have been developed for distribution-based mining of data for particular problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data [5].

In the perturbation approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations. For example, the classification technique uses a distribution-based analogue of single-attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be accordingly modified to work with the perturbation approach. This is because of the independent treatment of the different attributes by the perturbation approach.

This means that distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records. Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular for two main reasons:

Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast tool set of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

2.2 Randomized Response Techniques

The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information [10, 13]. Although information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items. Randomized Response technique was first introduced by Warner as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A, queries are sent to a group of people Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models:

Related-Question Model and Unrelated-Question Model have been proposed to solve this survey problem. In the Related-Question Model, instead of asking each respondent whether he/she has attribute A the interviewer asks each respondent two related questions, the answers to which are opposite to each other .When the randomization method is carried out, the data collection process consists of two steps. The first step is for the data providers torandomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm.

One main gain of the randomization method is that, it is relatively simple and does not require knowledge of the distribution of other records in the data. Therefore, the randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

2.3 Condensation Methodology

We introduce a condensation approach, [11] which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters .We refer to the technique as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. The constraints on the clusters are defined in terms of the sizes of the clusters which are chosen in a way so as to preserve *k* anonymity. This method has a number of advantages over the perturbation model in terms of preserving privacy in an effective way. In addition, since the approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data. Furthermore, the use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data. In contrast, when the data is constructed with the use of generalizations or suppressions, we need to redesign data mining algorithms to work effectively with incomplete or partially certain data. It can also be effectively used in situations with dynamic data updates such as the data stream problem. We discuss a condensation approach for data mining. This approach uses a methodology which condenses the data into multiple groups of predefined size, for each group, certain statistics are maintained. Each group has a size at least k, which is referred to as the level of that privacy-preserving approach. The greater the level, the greater the amount of privacy. At the same time, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. We use the statistics from each group in order to generate the corresponding pseudo-data.

2.4. Cryptographic Methodology

This branch became hugely popular [6] for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [7,10] has pointed that cryptography does not protect

the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

3. Assessment of Privacy Preserving Algorithms

3.1 Heuristic-Based Techniques

A number of techniques have been developed for a number of data mining techniques like classification, association rule discovery and clustering, based on the premise that selective data modification or sanitization is an NP-Hard problem, and for this reason, heuristics can be used to address the complexity issues.

3.1.1 Centralized Data Perturbation-Based Association Rule Confusion

A formal proof that the optimal sanitization is an NP Hard problem for the hiding of sensitive large item sets in the context of association rules discovery, have been given in [4].

3.1.2 Centralized Data Blocking-Based Association Rule Confusion

One of the data modification approaches which have been used for association rule confusion is data blocking [6]. The approach of blocking is implemented by replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. An approach which applies blocking to the association rule confusion has been presented in [22]. The introduction of this new special value in the dataset imposes some changes on the definition of the support and confidence of an association rule. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lies below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Notice that for an algorithm used for rule confusion in such a case, both 1-values and 0-values should be mapped to question marks in an interleaved fashion; otherwise, the origin of the question marks will be obvious. An extension of this work with a detailed discussion on how effective is this approach on reconstructing the confused rules, can be found in [14].

3.1.3 Centralized Data Blocking-Based Classification Rule Confusion

The work in [5] provides a new framework combining classification rule analysis and parsimonious downgrading. Notice here, that in the classification rule framework, the data administrator has as a goal to block values for the class label. By doing this, the receiver of the information, will be unable to build informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information from a secure environment (it is referred to as High) to a public one (it is referred to as Low), given the existence of inference channels. In parsimonious downgrading a cost measure is assigned to the potential downgraded information that it is not sent to Low. The main goal to be accomplished in this work is to find out whether the loss of functionality associated with not downgrading the data, is worth the extra confidentiality. Classification rules, and in particular decision trees are used in the parsimonious downgrading context in analyzing the potential inference channels in the data that needs to be downgraded. The technique used for downgrading is the creation of the so called parametric base set. In particular, a parameter θ, $0 \leq \theta \leq 1$ is placed instead of the value that is blocked. The parameter represents a probability for one of the possible values that the attribute can get. The value of the initial entropy before the blocking and the value of the entropy after the blocking is calculated. The difference in the values of the entropy is compared to the decrease in the confidence of the rules generated from the decision tree in order to decide whether the increased security is worth the reduced utility of the data the Low will receive. In [17] the authors presented the design of a software system, the Rational Down grader that is based on the parsimonious downgrading idea. The system is composed of a knowledge-based decision maker, to determine the rules that may be inferred, a "guard" to measure the amount of leaked information, and a parsimonious down grader to modify the initial down-grading decisions. The algorithm used to downgrade the data finds which rules from those induced from the decision tree induction, are needed to classify the private data. Any data that do not support the rules found in this way, are excluded from downgrading along with all the attributes that are not represented in the rules clauses. From the remaining data, the algorithm should decide which values to transforming to missing values. This is done in order to optimize the rule confusion. The "guard" system determines the acceptable level of rule confusion.

3.2 Cryptography-Based Techniques

A number of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature. Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. In particular, an SMS problem deals with computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation, and that no more

information is revealed to a participant in the computation than that's participant's input and output.

3.2.1 Vertically Partitioned Distributed Data Secure Association Rule Mining

Mining private association rules from vertically partitioned data, where the items are distributed and each item set is split between sites, can be done by finding the support count of an item set. If the support count of such an item set can be securely computed, then we can check if the support is greater than the threshold, and decide whether the item set is frequent. The key element for computing the support count of an item set is to compute the scalar product of the vectors representing the sub-item sets in the parties. Thus, if the scalar product can be securely computed, the support count can also be computed. The algorithm that computes the scalar product, as an algebraic solution that hides true values by placing them in equations masked with random values, is described in [13].The security of the scalar product protocol is based on the inability of either side to solve *k* equations in more than *k* unknowns. Some of the unknowns are randomly chosen, and can safely be assumed as private. A similar approach has been proposed in [14].Another way for computing the support count is by using the secure size of set intersection method described in [8].

3.2.2 Privacy Preserving Clustering

An algorithm for secure clustering by using the Expectation-Maximization algorithm is presented in [8]. The algorithm proposed is an iterative algorithm that makes use of the secure sum SMC protocol.

3.3 Reconstruction-Based Techniques

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining.

3.3.1 Reconstruction-Based Techniques for Numerical Data

The work presented in [3] addresses the problem of building a decision tree classifier from training data in which the values of individual records have been perturbed. While it is not possible to accurately estimate original values in individual data records, the authors propose a reconstruction procedure to accurately estimate the distribution of original data values. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. For the distortion of values, the authors have considered a discretization approach and a value distortion approach. For reconstructing the original distribution, they have considered a Bayesian approach and they proposed three algorithms for building accurate decision trees that rely on reconstructed distributions. The work presented in [2, 7, 13, 17] proposes an improvement over the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm or distribution reconstruction. More specifically, the authors prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They also show that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. It is also shown, that the privacy estimates of [3] had to be lowered when the additional knowledge that the miner obtains from there constructed aggregate distribution was included in the problem formulation.

3.3.2 Reconstruction-Based Techniques for Binary and Categorical Data

The work presented in [18] and [13] deal with binary and categorical data in the context of association rule mining. Both papers consider randomization techniques that offer privacy while they maintain high utility for the data set SIGMOD.

4. Distributed Privacy Preserving Data Mining

The key goal in most distributed methods for privacy-preserving data mining (PPDM) is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which has the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

4.1 Distributed algorithms over horizontally partitioned data sets

In horizontally partitioned data sets, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. Many of these techniques use specialized versions of the general methods discussed in for various problems. The work in discusses the construction of a popular decision tree induction method called ID3with the use of approximations of the best splitting attributes. Subsequently, a variety of classifiers have been generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier and the SVM Classifier with nonlinear kernels. An extreme solution for the horizontally partitioned case is discussed in [8], in which privacy preserving classification is performed in a fully distributed setting, where each customer has private access

to only their own record. A host of other data mining applications have been generalized to the problem of horizontally partitioned data sets [10]. These include the applications of association rule mining, clustering, and collaborative filtering.

4.2 Distributed algorithms over vertical partitioned data sets

For the vertically partitioned case, many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms. For example, the methods in [8] discuss how to use to scalar dot product computation for frequent item set counting. The process of counting can also be achieved by using the secure size of set intersection as described in [9]. Another method for association rule mining uses the secure scalar product over the vertical bit representation of item set inclusion in transactions, in order to compute the frequency of the corresponding item sets. This key step is applied repeatedly within the framework of a roll up procedure of item set counting. It has been shown that this approach is quite effective in practice. The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees, SVM Classification, Naïve Bayes Classifier, and k means clustering.

5. Evaluation of Privacy Preserving Algorithms

An important aspect in the development and assessment of algorithms and tools, for privacy preserving data mining is the identification of suitable evaluation criteria and the development of related benchmarks. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better that another one on specific criteria, such as performance and/or data utility. It is thus important to provide users with a set of metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are interested in optimizing. A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms is given below:

•the *performance* of the proposed algorithms in terms of time requirements, that is the time needed by each algorithm to hide a specified set of sensitive information;

•the *data utility* after the application of the privacy preserving technique, which is equivalent with the minimization of the information loss or else the loss in the functionality of the data:

•the *level of uncertainty* with which the sensitive information that have been hidden can still be predicted;

•the *resistance* accomplished by the privacy algorithms to different data mining techniques.

5.1 Performance of the proposed algorithms

A first approach in the assessment of the time requirements of a privacy preserving algorithm is to evaluate the computational cost. In this case, it is straightforward that an algorithm having a $O(n2)$ polynomial complexity is more efficient than another one with $O(en)$ exponential complexity. An alternative approach would be to evaluate the time requirements in terms of the average number of operations, needed to reduce the frequency of appearance of specific sensitive information below a specified threshold [5, 8, 12]. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a fast comparison among different algorithms. The *communication cost* incurred during the exchange of information among a number of collaborating sites, should also be considered. It is imperative that this cost must be kept to a minimum for a distributed privacy preserving data mining algorithm.

6. Conclusion

In this paper we have presented taxonomy of privacy preserving data mining approaches. With the development of data analysis and processing technique, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. While all the earlier procedures are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. In distributed privacy preserving data mining areas, efficiency is an essential issue, we should try to progress more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost. Privacy and Accuracy is a pair of contradiction; improving one usually incurs a cost in the other.

References

[1] Murat Kantarcioglou and Chris Clifton, *Privacy-preserving distributed mining of association rules on horizontally partitioned data*, In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery (2002), 24–31.

[2] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining*, In Advances in Cryptology - CRYPTO 2000 (2000), 36–54.

[3] P.Samarati,(2001). Protecting respondent's privacy in micro data release. In IEEE Transaction on knowledge and Data Engineering,pp.010-027.

[4] L. Sweeney, (2002)."k-anonymity: a model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.

[5] Agrawal, R. and Srikant, R, (2000)."Privacy-preserving data mining ".In Proc. SIGMOD00, pp. 439-450.

[6] Evfimievski, A.Srikant, R.Agrawal, and GehrkeJ(2002),"Privacy preserving mining of association rules". In Proc.KDD02, pp. 217-228.

[7] Hong, J.I. and J.A. Landay,(2004).Architecture for Privacy Sensitive Ubiquitous Computing", In Mobisys04, pp. 177- 189.

[8] Laur, H. Lipmaa, and T. Mieli' ainen,(2006)."Cryptographically private support vector machines".In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624.
[9] Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, (2005) "Template based privacy preservation in classification problems", In ICDM, pp. 466-473.
[10] Yang Z., ZhongS.Wright R.(2006) Privacy-Preserving Classification of Customer Data without Loss of Accuracy. SDM Conference, 603-610.
[11] Clifton C. Kantarcioglou M., Lin X., Zhu M.(2002).Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations, 4(2).
[12] M. Kantarcioglu and C. Clifton, (2002).“Privacy-preserving distributed mining of association rules on horizontally partitioned data”, In Proc.of DKMD’02
[13] Nabil Adam and John C. Wortmann, *Security-Control Methods for Statistical Databases: AComparison Study*, ACM Computing Surveys 21 (1989), no. 4, 515–556.
[14] DakshiAgrawal and Charu C. Aggarwal, *On the design and quantification of privacy preservingdata mining algorithms*, In Proceedings of the 20th ACMSymposium on Principles of Database Systems (2001), 247–255.
[15] RakeshAgrawal and RamakrishnanSrikant,*Privacy-preserving data mining*, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450.
[16] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, *Disclosure Limitation of Sensitive Rules*, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999),45–52.
[17] LiWu Chang and Ira S. Moskowitz, *Parsimoniousdowngrading and decision trees applied to the inference problem*, In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82–89.
[18] Chris Clifton and Donald Marks, *Security and privacy implications of data mining*, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

About the Authors:

Md. Riyazuddinis working as an Assistant Professor, Department of IT at Muffakham Jah College of Engineering and Technology (MJCET), Hyderabad, India. He has received M.Tech.(CSE) from JNTUH, Hyderabad. He has 8 years’ experience in teaching and published 4 Research papers in various International Journals. His main research interests are Data Mining, Cloud computing Computer Networks, Information Security and Software Engineering.

Dr. V V S SSBalaram is working as Professor and HOD in the Department of Information Technology at Sreenidhi Institute of Science and Technology (SNIST),Hyderabad, India. He has 17 years of teachingexperience. He did his M.Tech from Andhra University and Ph.D from Osmania University. He has been published and presented good number of research and technical articles in International, National Journals and International, National Conferences. His main research interests are Network Security and Cryptography, Data warehousing and Mining, Operating Systems, Distributed Operating Systems and Computer Graphics.

Md. Afrozeis working as an Assistant Professor , Department of IT at Muffakham Jah College of Engineering and Technology (MJCET), Hyderabad, India. He has received M.Tech.(CS) from JNTUH, Hyderabad. He has 8years experience in teaching and published 3 Research papers in various International Journals . His main research interests are Data Warehousing, Data Mining, Object Oriented Programming, Operating Systems, Web Technology, Object Oriented Modeling and Database Management Systems.

Md. JaffarSadiqis working as an Associate Professor , Department of IT at Sreenidhi Institute of Science and Technology (SNIST), Hyderabad, India. He has received M.Tech.(CSE) from JNTUH, Hyderabad. He has 8years experience in teaching. His main research interests are Data Mining, Image Processing, Cloud Computing, Computer Networks and Software Engineering.

M.D.Zuber working as an Assistant Professor Department of ComputerScience and Engineering at Hi-Point college of Engineering and Technology, Moinabad, Hyderabad, India. He has received M.Tech in Software Engineering from JNTUH, Hyderabad, AndhraPradesh, India. He has 5 years of teaching experience. His research interests are ImageProcessing, Data Mining,Computer Networks, and High PerformanceComputing.