An Efficient Data Mining Method for Clustering on Privacy Preserving Concept

R. Sasikala[1] T. Bhuvaneswari[2]
[1]Assistant Professor [2]M.Phil
[2]Department of Computer Science and Engineering
[1,2]Sankara College of Commerce and Science

Abstract— Privacy preserving data mining has become increasingly popular because it allows sharing of private sensitive data for analysis purposes. The concept of privacy preserving data mining has been proposed in response to these privacy concerns. The main goal of this research work has introduced a new k-Anonymity algorithm which is capable of transforming a non anonymous data set into a k-Anonymity data set. K-Anonymity model is thus to transform a table so that no one can make high-probability associations between records in the table and the corresponding entities. In order to achieve this goal, the K-Anonymity model requires that any record in a table be indistinguishable from at least (k−1) other records with respect to the pre-determined quasi-identifier. Finally the modified dataset is used for clustering.

Key words: Data mining, privacy preserving, Clustering

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers.

The task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

A. The Foundations of Data Mining:

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

In the evolution from business data to business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

The most commonly used techniques in data mining are:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.

All rights reserved by *www.ijsrd.com*

– Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms.An Architecture for Data Mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large datawarehouse.

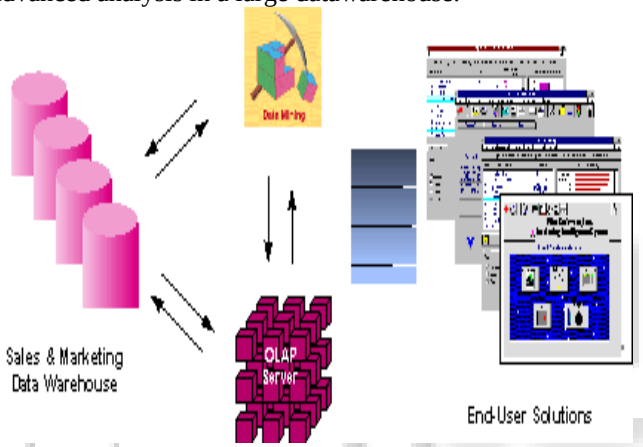

Fig. 1: Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

II. Review of Literature

The randomization method: The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records [2, 5]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions. We will describe the randomization technique in greater detail in a later section.

The k-anonymity model and l-diversity: The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, we reduce the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. The l-diversity model was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. To do so, the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme.

Distributed privacy preservation: In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of a variety of protocols. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data.

Downgrading Application Effectiveness: In many cases, even though the data may not be available, the output of applications such as association rule mining, classification or query processing may result in violations of privacy. This has lead to research in downgrading the effectiveness of applications by either data or application modifications. Some examples of such techniques include association rule hiding , classifier downgrading , and query auditing

All rights reserved by *www.ijsrd.com*

III. PPDM TECHNIQUES

A. Privacy Requirements:

Privacy is an important concern while disclosing various categories of electronic data including business data and medical data for data mining. Privacy can be interpreted in two ways. For instance, privacy is so crucial with respect to medical data, since it contains sensitive information like type of disease. Especially for doing medical data mining the original data should be available for making accurate predictions otherwise lead to impractical solutions. Any kind of disclosure related to the person- specific information leads to many problems including ethical issues. Therefore extra care should be taken to protect privacy of individuals before publishing such data. On the other hand, the privacy can be interpreted as preventing unwanted disclosure of information while performing data mining on aggregate results. Thus, privacy can be addressed at various levels in the process of data mining. For entire database security both privacy and security measures are needed. For better understanding of the concept of privacy, we would like to distinguish between the two related issues security and privacy according to HealthCare data. And the remaining subsections provide an introduction to privacy issues and privacy policies.

B. Security Vs Privacy:

Even though the two terms, security and privacy are synonymously used, these can be treated as two related, but separate issues: i) Security is defined as the mechanism for protecting the entire HealthCare data including the ability to control access to patient information, safeguard from unauthorized disclosure, alteration, loss or destruction of patient information. Security is typically accomplished through operational and technical controls. The three fundamental security goals are Confidentiality, Integrity and Availability. And ii) Privacy is a more specific term which is defined as the right of an individual to keep his/her individual health information from being disclosed. Privacy is typically accomplished though policies and procedures. With this understanding it is clear that security is necessary, but not sufficient for addressing privacy. Today several known PPDM techniques are available and these are extensively studied in literatures.

C. Privacy Issues

The privacy issue varies according to the data in use and the context it is used. But, the most important issue is how to provide privacy while preserving information (that is without loss of information).The methods like attribute removal, data hiding, and data compression can be applied on the data set to provide privacy, but will lead to information loss. Another important issue is regarding the computational overhead. Complex procedures like cryptographic techniques create additional overhead both technical and computational. The main parameter that affects the feasibility of implementing a secure protocol based on the generic constructions is the size of the best combinatorial circuit that computes the function that is evaluated.

For a distributed environment, when the number of parties becomes bigger, the communication and computational cost grow exponentially. The PPDM algorithm that addresses all these issues is still a myth. Even though no such generic solutions are available to address all privacy issues, some research has focused on finding efficient protocols for specific problems that balance privacy, data utility and computational feasibility.

IV. METHODOLOGY

The proposed methodology is used for analyzing K-anonymization of data using partitional hierarchical approach. The framework for the work is as given in Fig 4. The framework consists of two phases. In the first phase the k-anonymization of data is done, in the second phase k-means clustering is applied to the k-anonymized dataset. The methodology consists of the following steps:

1) Dataset
2) Quasi-identifiers
3) Suppression
4) Clustering
   - Clustering for original dataset
   - Clustering for anonymized dataset
5) Performance factors
   - Clustering accuracy for original and anonymized dataset
   - Time Complexity
   - Information loss

1. Consider the given dataset
2. Select the quasi-identifier attributes from the dataset.
3. Based on the threshold value k, suppress the quasi-identifier attributes
4. From the quasi-identifiers select the categorical attributes. The values of this categorical attributes are base items (n).
5. Convert the categorical attributes in to numerical attributes.
6. Find out the similarity distance using

$$D_{xy} = \frac{|m(X,Y)|}{|m(X)|+|m(Y)|-|m(X,Y)|}$$

7. Apply the HAC clustering algorithm to group the dataset.
8. Calculate the centroid of every formed cluster and add every categorical items as additional attributes of the centroid.
9. Now apply the k-means clustering algorithm.
10. Compare the k-means clustering performance of the original dataset and modified dataset.

A. Dataset:

The Adult dataset is downloaded from UC Irvine Machine Learning Repository. The Donor of the dataset is Ronny Kohavi and Barry Becker. This dataset contains the census data and has become a commonly used benchmark for K-Anonymity. The Adult dataset consists of fifteen fields with six continuous attributes and eight categorical attributes. The class attribute is income level with, two possible values, <= 50K or > 50K. The Adult data contains about 32,561 records totally.

B. Quasi-identifiers:

Quasi-identifiers are set of features whose associated values may be useful for linking with another data set to reidentify

All rights reserved by www.ijsrd.com

the entity that is the subject of the data For K-Anonymization, {age, work class, education, marital status, occupation, race, gender, and native country} are considered as the quasi-identifiers. Among these, age and education were treated as numeric attributes while the other six attributes were treated as categorical attributes.

C. Suppression:

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "?," indicating that any value can be placed instead. Based on the threshold value k, suppress the quasi-identifier attributes. The threshold value for suppression of quasi-identifiers is defined as k=2.

D. Clustering:

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms.

E. Categorical to Numerical Conversion:

In order to explore the relationships among categorical items, the idea of co-occurrence is applied. The basic assumption of co-occurrence is that if two items always show up in one object together, there will be a strong similarity between them. When a pair of categorical items has a higher similarity, they shall be assigned closer numeric values. The proposed algorithm produces pure numeric attributes.

The first step in the proposed approach is to read the input data and normalize the numeric attributes' value into the range of zero and one. The goal of this process is to avoid certain attributes with a large range of values will dominate the results of clustering. Additionally, a categorical attribute A with most number of items is selected to be the base attribute, and the items appearing in base attribute are defined as base items. This strategy is to ensure that a non-base item can map to multiple base items. If an attribute with fewer items is selected as the base attribute, the probability of mapping several nonbased items to the same based items will be higher. In such a case, it may make different categorical items get the same numeric value.

After the based attribute is defined, counting the frequency of co-occurrence among categorical items will be operated in this step. A matrix M with n columns and n rows is used to store this information, where n is the number of categorical items which represents the appearance of co-occurrence between the base items.

Example of a sample dataset

Attribute W	Attribute X	Attribute Y	Attribute Z
A	C	0.1	0.1
A	C	0.3	0.9
A	D	0.8	0.8
B	D	0.9	0.2
B	C	0.2	0.8
B	E	0.6	0.9
A	D	0.7	0.1

Table 1:

$$M = \begin{pmatrix} 4 & 0 & 2 & 2 & 0 \\ 0 & 3 & 1 & 1 & 1 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Since the frequencies of co-occurrence between base items and other categorical items is available by retrieving the elements in matrix M, the similarity between them can be calculated by adopting following equation.

$$D_{xy} = \frac{|m(X,Y)|}{|m(X)| + |m(Y)| - |m(X,Y)|}$$

Where X represents the event that item x appears in the set of objects; Y represents the event that item y appears in the set of objects; m(X) is the set of objects containing item x; m(X, Y) is the set of objects containing both item x and y. In the above equation, when two items always show up together in objects, the similarity between them will be one. If two items never appear together, it will get zero for the similarity measure. The higher value of Dxy means the more similar between item x and item y. However, only the values of Dxy larger than a threshold will be recorded, or zero will be assigned.

All attributes in data set will contain only numeric value at this moment, the existing distance based clustering algorithms can be applied without pain. HAC (Hierarchical Agglomerative Clustering) is a widely used hierarchical clustering algorithm. The major difference is the applied similarity criteria. The HAC algorithm takes numeric data as the input and generates the hierarchical partitions as the output. Therefore it is applied in first clustering step to group data into subsets. In HAC, initially each object is considered as a cluster. Then by merging the closest clusters iteratively until the termination condition is reached, or the whole hierarchy is generated.

F. HAC Algorithm:

1) Calculate the distance between every two objects.
2) View each object as an individual cluster.
3) Merge the closest two clusters.
4) Update the distance between clusters.
5) Repeat 3-4 until reaching a stopping criterion or generating the whole hierarchy.

The k-means algorithm takes numeric data as input, and generates crispy partitions (i.e., every object only belongs to one cluster) as the output. It is one of the most popularly used clustering algorithms in the research community. It has been shown to be a robust clustering method in practice. Therefore, the k-means algorithm is applied in second

All rights reserved by www.ijsrd.com

clustering step to cluster data sets. K-means starts by randomly selecting or by specifically picking k objects as the centroids of k clusters. Then k-means iteratively assigns the objects to the closest centroid based on the distance measure, and updates the mean of objects in this cluster as the new centroid until reaching a stopping criterion. This stopping criterion is based on either non-changing clusters or a predefined number of iterations.

G. K-Means Algorithm:

1) Select first k objects randomly as the centroid of each cluster.
2) Assign each object to the closest cluster based on Euclidean distance or cosine similarity.
3) Update the centroid of each cluster.
4) Repeat steps 2-3 until stopping criterion is reached.

V. RESULTS AND DISCUSSION

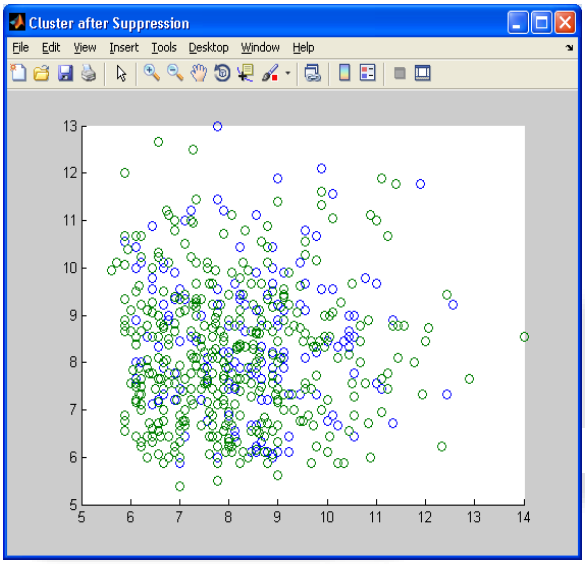

Fig. 2: Clustering results

The transformation is aimed to achieve a predicate performance of a clustering algorithm trained on a transformed data set as similar as possible to the performance of a cluster trained on the original data set. In this method we are applied a k-means algorithm with bisection method for predict a cluster. So after suppression the data set we are analysis the following factors such as clustering accuracy for original and suppression data set, Time Complexity and information loss. For this paper we are taken an adult dataset to computing the various techniques like clustering and bisection methods for predicating the cluster. So, here additionally predict the suppression for given data set, at last we are find the cluster before suppression and after suppression, time complexity , accuracy find before suppression and after suppression accuracy and information loss. Data points are given by the user during the execution of the program. For different input data points, the algorithm gives different outputs.

VI. CONCLUSION AND FUTURE WORK

K-anonymity has recently been investigated as an interesting approach to protect microdata undergoing public or semi-public release from linking attacks. Using data mining techniques as a basis for k-anonymization has two major benefits, which arise from the fact that different data mining techniques consider different representations of data. In that such anonymization algorithms are optimized to preserve specific data patterns according to the underlying data mining technique. This work proposes an clustering algorithm for K-Anonymized data. Clustering algorithms has been widely applied to various domains to explore the hidden and useful patterns inside data. Because the most collected data in real world contain both categorical and numeric attributes, the traditional clustering algorithm cannot handle this kind of data effectively. Therefore, in this work we propose a new approach to explore the relationships among categorical items and convert them into numeric values. Moreover, in order to overcome the weaknesses of k-means clustering algorithm, a two-step method integrating hierarchical and partitioning clustering algorithms is introduced.

REFERENCES

[1] Agrawal and Srikant, "Privacy Preserving Data mining", Proceedings of the ACM SIGMOD International Conference on Management of data, 2000.
[2] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data", Data & Knowledge Engineering 65 (2008) 5–21.
[3] Samarati P, "Protecting respondent's privacy in Microdata release", IEEE Transactions on Knowledge and Data Engineering, 13:1010–1027
[4] Sweeney L, "k-anonymity: A model for protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570.
[5] Tiancheng Li, Ninghui Li, "Towards Optimal k-anonymization", Data & Knowledge Engineering, 2008 Elsevier.
[6] Geetha Jagannathan, Rebecca N. Wright, "Privacy-Preserving Imputation of Missing Data", Data & Knowledge Engineering, 2008 Elsevier.
[7] Jiang, Clifton and Kantarcıoğlu, "Transforming Semi-Honest Protocols to Ensure Accountability", Data & Knowledge Engineering, 2008 Elsevier.
[8] Bhavani Thuraisingham, "Privacy constraint processing in a privacy-enhanced database management system", Data & knowledge Engineering, 2005.
[9] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on Computer and Communication Technology, 2012.
[10] Jian Wang, Yongcheng Luo, Yan Zhao, Jiajin Le, "A Survey on Privacy Preserving Data Mining", First International Workshop on Database Technology and Applications, 2009.
[11] Benny Pinkas, "Cryptographic techniques for privacy preserving data mining", http//www.pinkas.net/PAPERS/sigkdd.pdf.

All rights reserved by *www.ijsrd.com*