ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Data Mining: Confidentiality of Data Preservation

Ms.Ifath Nazia Ghori
HoD,CS & IT,
Girls Community College,
Abu Arish, Jazan University,KSA.

Mr.Fahad Bin Awad Quraishi
P.G Scholar,
Dept of CSE,
Hi Point College of Engineering & Technology.

ABSTRACT:

Vast amount of detailed personal data is regularly collected and sharing of these data is proved to be advantageous for data mining applications. Such data include shopping habits of supermarkets, criminal records, medical history, credit records, finance records, school and college records etc .On one hand such data is an important asset to business group and governments for decision making by analyzing the same data.On the other hand privacy policy and other privacy concerns may prevent data owners from strictly sharing information for data analysis.

In order to share data while preserving privacy data owner must come up with a solution which achieves the dual goal of privacy preservation as well as accurate clustering result. Trying to give solution for this we have implemented vector quantization approach piecewise on the datasets which segmentizes each row of datasets and quantization approach is performed on each segment using "K" means which later are again united to form a transformed data set.Some tentative results are presented which tries to finds the optimum value of segment size and quantization parameter which gives optimum in the trade-off between clustering utility and data privacy in the input dataset.

Keywords:

Data Warehousing, Datamining, Data Security, Preserving Datamining, Quantization Approach, predictive Information, Ppdm, Ppdp, Segmentation, DBMS.

INTRODUCTION:

Data mining is the extraction of hidden predictive information from large databases, which helps businesses to make proactive, knowledge-driven decisions Data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The automated, prospective analysis offered by data mining to predict future trends and behaviors moves beyond the analyses of past events provided by retrospective tools. Data mining answers business questions that traditionally have been too time-consuming or complex to answer. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information.

Long process of research and product development_ evolved.datamining. 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:

DataminingAlgorithms	Powerful Multiprocessor Computer	Massive Data Collection

The Scope of Data Mining :

Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

• Automated prediction of trends and behaviors:

Data mining automates the process of finding predictive information in large databases.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

• Automated discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. Databases can be larger in both depth and breadth:

• More rows:

Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

• More columns:

Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

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 process 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.

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.

Applications of DataMining:

There is a rapidly growing body of successfulapplications in a wide range of areas as diverse as:Analysis of organic compounds, automatic abstracting,credit card fraud detection, financial forecasting, medicaldiagnosis etc. Some examples of applications (potential oractual) are:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

A supermarket chain mines its customer transactions-data to optimize targeting of high value customers

i.A credit card company can use its data warehouse of-customer transactions for fraud detection

ii.A major hotel chain can use survey databases toidentify attributes of a 'high-value' prospect.

Applications can be divided into four main types:

Classification	Numerical Prediction
Association	Clustering

Data mining using labeled data (specially designate-dattribute) is called supervised learning. Classification andnumerical prediction applications falls in supervisedlearning. Data mining which uses unlabeled data is-termed as unsupervised learning and association and-clustering falls in this category.

DATAMINING AND PRIVACY:

Datamining deals with large database which can containsensitive information. It requires data preparation whichcan uncover information or patterns which may-compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has-increased the disclosure risks of sensitive data. Acommon way for this to occur is through dataaggregation. Data aggregation is when the data areaccrued, possibly from various sources, and put togetherso that they can be analyzed.

This is not data mining perse, but a result of the preparation of data before and forthe purposes of the analysis. The threat to an individual'sprivacy comes into play when the data, once compiled,cause the data miner, or anyone who has access to thenewly compiled data set, to be able to identify specificindividuals, especially when originally the data wereanonymous.

What data mining causes is social and ethical problem byrevealing the data which should require privacy?Providing security to sensitive data against unauthorizedaccess has been a long term goal for the database securityresearch community and for the government statisticalagencies.

Hence, the security issue has become, recently, amuch more important area of research in data mining. Therefore, in recent years, privacy-preserving data mining has been studied extensively.

PRIVACY-PRESERVINGDATAMINING:

The recent work on PPDM has studied novel data miningtechniques that do not require accessing sensitiveinformation. The general idea of PPDM is to allow datamining from a modified version of the data that containsno sensitive information.

CENTRALIZEDMODEL:

In the centralized model, all data are owned by a singledata publisher. The key issues are how to modify the dataand how to recover the data mining result from themodified data. Answers often depend on data miningoperations and algorithms. One common technique israndomization, by introducing random noise andswapping values in the data. The randomized datapreserves aggregate properties (such as means andcorrelations) in the collection of records, but has little usewhen each record is examined individually.

Anothercommon technique is encryption. The data publishertransforms the original data into an encrypted form fordata mining at an external party. Since the data miningresults are in the encrypted form and since the datapublisher is the only one who can decrypt the results, thisapproach is applicable only if the data publisher himselfis the user of data mining results.

DISTRIBUTEDMODEL:

In the distributed model, multiple data publishers wantto conduct a computation based on their private inputs,but no data publisher is willing to disclose its own outputto anybody else. The privacy issue here is how to conductsuch a computation while preserving the privacy of theinputs. This problem is known as the Secure MultipartyComputation (SMC) problem. The aim of SMC is toenable multiple parties to carry out distributedcomputing tasks in a secure manner with the assumptionthat some attackers, who possibly are the participatingparties themselves, want to obtain extra information otherthan the output. SMC has two major requirements,privacy and correctness.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

The privacy requirement statesthat parties should learn their output and nothing elseduring and after the SMC process. The correctnessrequirement states that each party should receive itscorrect output without alteration by the attackers.Extensive research hasbeen conducted on secureprotocols for data mining tasks including association rulemining, classification analysis, and clustering analysis.Refer to for surveys on this distributed model of PPDM.

COMPARINGPPDP ANDPPDM:

In many real life applications, the data publisher wants topublishsome data, but has little or no interest in datamining results and algorithms. For example, a hospitalmay publish the patient data to a drug research institute;although willing to contribute its data to drug research,the hospital is not interested in and has no expertise indata mining algorithms because drug research is not itsnormal business. This privacy-preserving data publishing(PPDP) scenario differs from PPDM in several majorways. PPDP focuses on the data, not data mining results;therefore, published records should be meaningful whenexamined individually.

This implies that randomizationand encryption are inapplicable. PPDP seeks toanonymize the data by hiding the identity of individuals,not hiding sensitive data. The anonymized data isexpected to be analyzed by traditional data miningtechniques; therefore, no new data mining techniques areneeded. We did not intend to dismiss the contribution ofthe randomization and encryption approaches. They areeffective anonymization methods if the data records willnot be examined individually.

CONFIDENCEBOUNDING:

A solution is possible. In particular, we iteratively disclosedomain values in a top-down manner by suppressing all domain values. In eachiteration, we disclose the suppressed domain value tomaximize some criterion taking into account bothinformation gained and privacy lost. We evaluate thismethod on real life data sets. Several features make thisapproach practically useful:

No taxonomy required. Suppression replaces a domainvalue with?without requiring a taxonomy of values.

This is a useful feature becausemost data do not have an associated taxonomy, though taxonomies may exist incertain specialized domains.

• Preserving the truthfulness of values. The special value?represents the union," a less precise but truthful representation, of suppresseddomain values. This truthfulness
is useful for reasoning and explaining the classificationmodel.

• Subjective notion of privacy. The data publisher has theexibility to her own notion of privacy using templates forsensitive inferences. Excient computation. It operates on simple but effectivedata structures to reduce the need for accessing raw data records.

• Anytime solution. At any time, the user (the datapublisher) can terminate the com putation and have a table satisfying the privacy goal.

• Extendibility. Though we focus on categorical attributesand classification analysis, this work can be easily extended to continuous attributesand other information utilitycriteria.

CONCLUSIONS:

Due to the wide use of the Internet and the trends ofenterprise integration, one-stop service, simultaneous cooperation and competition, andoutsourcing in both public and private sectors, datapublishing has become a daily and routine activity ofindividuals, companies, organizations, governmentagencies. Privacy-preserving data publishing is apromising approach for data publishing withoutcompromising individual privacy or disclosing sensitiveinformation. In this thesis, we studied different types oflinking attacks in the data publishing scenarios of singlerelease, sequential release, and secure data integration .Our contributions can be summarized.as.follows:

» Preserving Privacy and Information.We considered theproblem of protecting individual privacy while releasing person-specific data forclassification modeling. We chose classification analysis as the informationrequirement because the data quality and usefulness can be objectively measured. Ourproposed framework can easily adopt other information requirement with a different selectioncriterion.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

» A Unified Privacy Notion.We demeaned a new privacynotion, called privacy template in the form ofhX; Y; ki,that unique anonymity template and confidentialitytemplate.This unified notion is applicable to all data-publishing scenarios studied in this thesis.

» A Framework of Anonymization Algorithm.Despite thedata publishing scenarios are very different; we presented a framework ofanonymization algorithm, Top-Down Renement (TDR), to iteratively specialize the data from ageneral state into a special state, guided by maximizing the information utility andminimizing the privacy specicity. This top-down approach serves a natural andefficient structure for handling categorical and continuous attributes and multipleprivacy templates. Experiments suggested that our TDR framework effectively preservesboth information utility and individual privacy and scales well for large data sets indifferent data publishing scenarios.

» Extended Data Publishing Scenarios.Most existing worksconsidered the simplest data publishing scenario, that is,a single release from a single publisher. Such mechanismsare insufficient because they only protect the data up tothe release or the sapient. Therefore, we also extended theprivacy notion and anonymization framework to other real life data publishing scenarios, includingsequential release publishing and secure data integration.