PRIVACY PRESERVATION TECHNIQUES IN DATA MINING

Jharna Chopra[1], Sampada Satav[2]

[1] M.E Scholar, CTA, [2] Asst. Prof, CSE, SSCET, Bhilai, CG, India,
jharna.chopra@gmail.com, sampada.satav@gmail.com

Abstract

In this paper different privacy preservation techniques are compared. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples . For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier

Index Terms: *Data Mining, Privacy Preservation, Clustering, Classification Techniques, Naive Bayes.*

----------***----------

1. INTRODUCTION

Data mining is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human-understandable structure and involves database and data management, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structure, visualization and online updating.

Privacy has become an increasingly important issue in data mining [5]. Privacy concerns restrict the free flow of information. Privacy is one of the most important properties of information. The protection of sensible information has a relevant role Organization for legal a commercial do not want to reveal their private database and information. Even the individual does not want to reveal their personal data to other than those they give permission to. This implies that revealing of an instance to be classified may be equivalent to revealing secret and private information .The protection of sensible information has a relevant role .Privacy preserving data mining algorithms have been recently introduced with the aim of preventing the discovery of sensible information .

Privacy-Preserving Data Mining is developing models without seeing the data is receiving growing attention [4]. Privacy-preserving data mining considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed even to the party running the algorithm. There are two significance settings for privacy-preserving data mining In the first, the data is divided amongst two or more different parties, and the aim is to run a data mining algorithm on the union of the parties' databases without allowing any party to view anyone else's private data. In the second, some statistical data that is to be released (so that it can be used for research using statistics and/or data mining) may contain confidential data and so is first modified so that (a) the data does not compromise anyone's privacy, and (b) it is still possible to obtain meaningful results by running data mining algorithms on the modified data set [3].

This paper address issues related a privacy preserving data classification and its advantages .Data source collaborate to develop a global model but the data is not disclosed to others. Naïve Bayes classifier is used as baseline as it provides reasonable classification performance.

2. PRIVACY PERSERVATION TECHNIQUES

A). Privacy Preserving Association rule Learning

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. The goal of association rule learning is to find specific patterns that represent knowledge in generalized form without referring to particular data item. Because of this one might say that association rule learning only represents an indirect threat to privacy. However traditional methods require access to the data set in order to be able to find association rules. form without referring to particular data items.

B). Privacy preserving Clustering techniques

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. The goal in clustering is to partition data elements into clusters so that the similarity among elements belonging to the same clusters is high, and so that the similarity among elements from different clusters is low. In privacy preserving clustering a main goal is to find the clusters in the data without revealing the content of the data elements themselves. the data may be partitioned vertically and/or horizontally among the involved parties

Types of clustering methods:-

a. Partitioning Methods
b. Hierarchical Agglomerative (divisive) methods
c. Density based methods
d. Grid-based methods

C). Privacy Preserving Classification Techniques

It is one of the biggest challenges in data mining . It is a predictive modeling task with the specific aim of predicting the value of a single nominal variable based on the known values of other variables[1].Classification is the task of generalizing known structure to apply to new data. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. A classification task begins with a data set in which the class assignments are known. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

Types of Classification Models:-

a) Decision Tree
b) K-.Nearest Neighbors
c) Artificial Neural Network
d) Support Vector Machine
e) Naive Bayes

a). Decision Tree

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Internal nodes are represented as circles, whereas leaves are denoted as tri-angles.

b). K-Nearest Neighbors

The k-nearest neighbor's algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance - based learning, or we can say it is lazy learning. It can also be used for regression. The k-nearest neighbour algorithm is amongst the simplest of all machines –learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric; however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

c) Artificial Neural Network

Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of

predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons "). Network is then subjected to the process of "training." In the training phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting "network" developed in the process of "learning" represents a pattern detected in the data

d) Support Vector Machine

Support Vector Machines were first introduced to solve the pattern classification and regression problems by Vapnik and his colleagues. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyper-plane in that space, one which maximizes the margin between the two data sets .To calculate the margin, two parallel hyper -planes are constructed, one on each side of the separating hyper-plane, which are "pushed up against" the two data sets . A good separation is achieved by the hyper -plane that has the largest distance to the neighbouring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier. This hyper -plane is found by using the support -vectors and margins.

e). Naïve Bayesian

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations.

Table 2.1: Advantage and Disadvantage of different Classification Algorithm

Classification Algorithm	Advantage	Disadvantage
DECISION TABLE	**1)**Decision trees are self–explanatory and when compacted they are also easy to follow. This representation is considered as comprehensible. **2)**Decision trees can handle both nominal and numeric input attributes. **3)**Decision tree representation is rich enough to represent any discrete value classifier. **4)**Decision trees are capable of handling datasets that may have errors. **5)**Decision trees are capable of handling datasets that have missing Value	**1)** Most of the algorithms (like ID3 and C4.5) require that the target attribute will have only discrete values. **2)** As decision trees use the "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present.
K-.NEAREST NEIGHBOURS	**1)**Because the process is transparent, it is easy to implement and debug. **2)**In situations where an explanation of the output of the classifier is useful k-NN can be very effective if an analysis of the neighbors is useful as explanation.. **3)**There are some noise reduction techniques that work only for k-NN that can be effective in improving the accuracy of the classifier .	**1)**Because all the work is done at run-time-KNN can have poor run-time performance if the training set is large. **2)**k-NN is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification. This can be ameliorated by careful feature selection or feature weighting.
ARTIFICIAL.NEURAL NETWORK	**1)**Neural network can perform tasks that a linear program cannot. **2)**When an element of the neural network fails, it can continue without any problem by their parallel nature. **3)** A neural network learns and does not need to be reprogrammed. **4)**It can be implemented in any application and without any problem.	**1)**The neural network needs training to operate. **2)**The architecture of a neural Network is different from the architecture of microprocessors therefore needs to be emulated **3)**Requires high processing time for large neural networks.
SUPPORT VECTOR MACHINE	**1)**By introducing the kernel, SVMs gain flexibility in the choice of the form of the threshold separating solvent from insolvent companies, which needs not be linear and needs not have the same functional form for all data. **2)**SVM deliver a unique solution	**1)** Lack of transparency of results

CONCLUSIONS

The objective of our work is to provide a study of different privacy preservation techniques. The various advantages and disadvantages are listed on the table. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective

REFERENCES

[1] Murat kantarcioglu and Jaideep Vaidya .PrivacyPreserving Naive Bayes Classifier for HorizontallyPartitioned Data. Purdue University.

[2] Zhihua Wei, Hongyun Zhang ,Zhifei Zhang, Wen Li, and duoquian Miao. July 2011.A Naive Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Result.Shanghai ,China.

[3]Zhiqiang Yang, Sheng Zhong, Rebecca n.Wright Privacy-Preserving Classification of Customer Data without Loss of Accuracy.Rutgers University,Piscataway.

[4] Nidhi Bhatia and Kiran Jyoti . An Analysis of Heart Disease Prediction using Different Data Mining Techniques. International Journal of Engineering Research & Technology Vol.1 Issue 8,ISSN:2278-0181

[5] Emmanouil Magkos , Manolis Maragoudakis, Vassilis Chrissikopoulos and Stefanos Gritzalis.13 April 2009.Accurate and Large Scale Privacy –Preserving Data mining Using the Election Paradigm.

[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, " From Data Mining to Knowledge Discovery in Databases", Providence, Rhode Island July 27–31, 1997.

[7] Mrs. Bharati and M. Ramageri ,"Data mining Technique and Applications" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305, ISSN : 0976- 5166.

[8] Lior Rokach and Oded Maimon," Decision Tree ", DataMining and Knowledge Discovery Handbook.

[9] Megha Gupta and Naveen Aggarwal," Classification techniques Analysis", NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19-20 March 2010.

[10]Goldreich and Oded, "Foundations of Cryptography": Volume 2, Basic Applications. Vol. 2. Cambridge university press, 2004.

[11] Benny Pinkas," Cryptographic techniques for privacy preserving data mining", SIGKDD Explorations. Volume 4, Issue 2 - page 18.

[12] Jaideep Vaidya ,Murat Kantarcıo and ˇglu · Chris Clifton," Privacy-preserving Naïve classification", Received: 30 September 2005 / Revised: 15 March 2006 / Accepted: 25 July 2006 / Published online: 3 February 2007 © Springer-Verlag 2007.

[13] Rivest, R.; A. Shamir; L. Adleman (1978). "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". Communications of the ACM 21 (2):120–126. doi:10.1145/359340.359342.

[14] Tjen-Sien Lim, Wei-Yin Loh , and Yu-Shih.A Comparison of Prediction Accuracy ,Complexity ,and Training Time of Thirty-Three Old and New Classification Algorithm.Machine Learning ,40,203- 2299 2000.

[15] M.M.J. Stevens (June 2007). On Collisions for MD5.