International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

An Encryption Scheme for Privacy Preserving in Data Mining using different classification Algorithm

Anu Verma[1], Jyoti Arora[2]

[1]M. Tech Student, Desh Bhagat University

[2]Assistant Professor, Desh Bhagat University

Abstract: ***Data mining is the process of extracting useful information from the large amount of database at any time and at any place. The normalization is a process of organizing the data in database to avoid data redundancy, insertion anomaly, update anomaly & deletion anomaly. If a database design is not perfect, it may contain anomalies, which are like a bad dream for any database administrator. Managing a database with anomalies is next to impossible. In the proposed work, data set of normalization is analyzed. The analyzation process is done to remove the missing values from the database. Then an encryption function is developed. Different classifiers are implemented for accuracy of data.***

Keywords: normalization, encryption, security, accuracy

1. Introduction

1.1 Data Mining

Data mining is crucial for extracting and identifying useful information from a large amount of data that is why retailing companies operate databases in a long way, such that all transactions are stored in arranged order. A record-of-transaction database typically contains the transaction date and the products bought in the course of a given transaction. Usually, each record also contains shopper ID, particularly when the purchase was made using a credit card or a frequent buyer card.

1.2 Customer Relationship Management (CRM)

CRM is the core business strategy that integrates internal processes and functions of the organization, to create and deliver value to targeted customers at a profit. It is mainly grounded on high quality customer related data and enabled by information technology. CRM is an information industry term that helps an enterprise to manage customer relationships in an organized way and helps the company to provide better services to its customers.

1.3 Classification In Data Mining

Classification is one kind of predictive modeling. More specifically, classification is the process of assigning new objects to predefined categories or classes. Given a set of labeled records, we build a model such as a decision tree, and predict labels for future unlabeled records. Model building in the classification process is a supervised learning problem. Training examples are described in terms of (1) attributes, which can be categorical—i.e., unordered symbolic values or numeric; and (2) class label, which is also called the predicted or output attribute. If the latter is categorical, then we have a classification problem. If the latter is numeric, then we have a regression problem. The training examples are processed using some machine learning algorithm to build a decision function such as a decision tree to predict labels of new data.

1.4 Classification Approaches

1.4.1 Decision Trees

Decision trees are the best-known classification paradigm. A decision tree represents a set of classification rules in a tree form. Each root–leaf path corresponds to a rule of form $T_{i1} \wedge \ldots \wedge T_{il} \rightarrow (C = c)$, where c is the class value in the leaf and each T_{ij} is a Boolean-valued test on attribute A_{ij}. The earliest decision trees were constructed by human experts.

1.4.2 Bayesian Classifiers

In Bayesian networks, statistical dependencies are represented visually as a graph structure. The idea is that we take into account all information about conditional independencies and represent a minimal dependency structure of attributes. Each vertex in the graph corresponds to an attribute and the incoming edges define the set of attributes, on which it depends. The strength of dependencies is defined by conditional probabilities. For example, if A1 depends on attributes A2 and A3, the model has to define conditional probabilities P (A1|A2, A3) for all value combinations of A1, A2 and A3.

1.4.3 K-Nearest Neighbor Classifiers

K-nearest neighbor classifiers represent a totally different approach to classification. They do not build any explicit global model, but approximate it only locally and implicitly. The main idea is to classify a new object by examining the class values of the K most similar data points. The selected class can be either the most common class among the neighbors or a class distribution in the neighborhood. The only learning task in K-nearest neighbor classifiers is to select two important parameters: the number of neighbors K and distance metric d. An appropriate K value can be selected by trying different values and validating the results in a separate test set.

1.4.4 Support Vector Machines

Support vector machines (SVMs) are an ideal method, when the class boundaries are non-linear but there is too little data to learn complex non-linear models. It is enough to save the support vectors, i.e. data points which define the class boundaries.

The main advantage of SVMs is that they find always the global optimum, because there are no local optima in maximizing the margin. Another benefit is that the accuracy does not depend on the dimensionality of data and the system is very robust to over fitting. This is an important advantage, when the class boundary is non-linear. Most other classification paradigms produce too complex models for non-linear boundaries.

1.4.5 Linear Regression

Linear regression is actually not a classification method, but it works well, when all attributes are numeric. For example, passing a course depends on the student's points, and the points can be predicted by linear regression. In linear regression, it is assumed that the target attribute (e.g. total points) is a linear function of other mutually independent attributes.

2. Related Work

Manjari Anand et al. [1] stated CRM is a kind of implemented model for managing a company's interactions with their customers. CRM involves the customer classification to understand the behavior of the customer. There is a vital role of the data mining techniques for the classification. This paper presents the concept of one of the data mining technique ART for the customer classification for CRM.

Ms. Saranya, et al. [2] stated customer Relationship management (CRM) is a seriously considered issue in today's competitive corporate world. For a firm to maintain an intact relationship with its customers, the vast amount of data within a business enterprise can be refined, mined and analyzed. This paper proposes a Decision Support System for analyzing and retaining the customers in an Online shopping System using various data mining techniques. To achieve successful Customer relationship management, various mining algorithms for pattern extraction from data have been utilized. Managerial decision makers can make use of these patterns for raising the profit graph of the firm.

Kamal R. et al. [3] suggested the choice value and the testing process against the vigilance parameter, characteristic of ART Neural Network, is merged. Only, a single unique test is required to determine if a committed category node can represent the current input or not. Advantages of APT over ART are: 1-Avoid testing every committed category node before deciding to train a committed category node or a new node must be committed, 2-The vigilance parameter is fixed during training, and 3-The choice value parameter is eliminated.

Mohammed Al-Maolegi et al. [4] stated that there are several mining algorithms of association rules. One of the most popular algorithms is Apriori that is used to extract frequent item sets from large database and getting the association rule for discovering the knowledge. Based on this algorithm, this paper indicates the limitation of the original Apriori algorithm of wasting time for scanning the whole database searching on the frequent item sets, and presents an improvement on Apriori by reducing that wasted time depending on scanning only some transactions. The paper shows by experimental results with several groups of transactions, and with several values of minimum support that applied on the original Apriori and our implemented improved Apriori that our improved Apriori reduces the time consumed by 67.38% in comparison with the original Apriori, and makes the Apriori algorithm more efficient and less time consuming.

Meenakshi et al. [5] suggested that in data mining, classification is to accurately predict the target class for each case in the data. Decision tree algorithm is one of the commonly used classification algorithm to make induction learning based on examples. In this paper, they present the comparison of different classification techniques using WEKA. The aim of this paper is to investigate the performance of different classification methods on clinical data. The algorithm tested are Bayes Network, Navie bayes, Logistic, rule Jrip, and J48.

3. Problem Formulation

Data mining is the process of extracting useful information from the large amount of database at any time and at any place. The normalization is a process of organizing the data in database to avoid data redundancy, insertion anomaly, update anomaly & deletion anomaly. If a database design is not perfect, it may contain anomalies, which are like a bad dream for any database administrator. Managing a database with anomalies is next to impossible. In the proposed work, data set of normalization is analyzed. The analyzation process is done to remove the missing values from the database. Then an encryption function is developed. Different classifiers are implemented for accuracy of data.

4. Methodology

Step 1: In first step we analyze the data set for normalization. Normalization is creation of shifted and scaled versions of statistics where the intention is that these normalized values allow the comparison of corresponding normalized values for different datasets So that we can removed the missing values.

Step 2: In second step we develop an encryption function that converts originality of the data for security aspects. In Encryption we encode messages or information in such a way that only authorized parties can read it. Encryption does not of itself prevent interception, but denies the message content to the interceptor. Encryption is use for security purpose.

Step 3: In last step we implement different classifiers for the prediction of the accuracy of the original as well as the encrypted data. An algorithm that implements classification in a concrete implementation is known as a classifier. In this

step we find the accuracy between the original data and encrypted data.

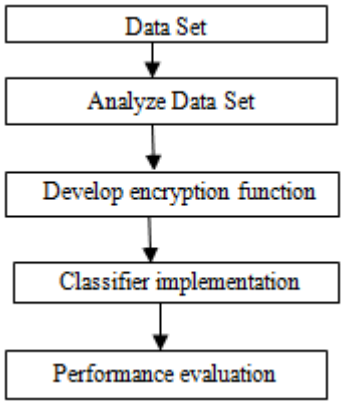

Figure 4.1: Flow of work

5. Results and Discussions

Table 5.1: Accuracy table for different classifier using original and encrypted dataset

Classifier	Original	Encrypted
Decision Table	71.48	75.09
JRip	72.40	73.56
J48	73.82	77.39
Random Forest	74.86	78.92

This table represents classification accuracy of different classifier based on original and encrypted dataset. Accuracy has been measure on the basis of correctly classified instances.

Table 5.2: Parameter table for encrypted dataset

Classifier	Precision	Recall	F-measure	TP rate	TN rate	Roc Area
Decision Table	74.3	75.1	74.5	75.1	36.7	78.2
JRip	72.3	73.6	72.4	73.6	41.3	67.6
J48	78.2	77.4	74.5	77.4	40.0	65.3
Random Forest	78.7	78.9	78.8	78.9	28.5	82.6

This table represents various performance evaluation parameters for data classification using encrypted dataset. On the basis of these parameters algorithm efficiency on dataset has been measured.

Table 5.3: Parameter table for original dataset

Classifier	Precision	Recall	F-measure	TP rate	TN rate	Roc Area
Decision Table	70.6	71.5	70.8	71.5	33.2	76.1
JRip	75.5	76	75.5	76	32.2	73.9
J48	73.5	73.8	73.6	73.8	32.7	65.3
Random Forest	74.4	74.9	74.5	74.9	32.5	81.5

This table represents various performance evaluation parameters for data classification using original dataset. On the basis of these parameters algorithm efficiency on dataset has been measured.

6. Conclusion

Data mining is the process for extraction of valuable information from raw information. Various classification, clustering and association based approaches have been used for extraction of value able information. In the purposed work security in classification is main concern. Due to lack of security in classification valuable information of any organization can be leaked that corresponds to various problems in an organization. In the purposed work diabetes dataset has been used for classification. Diabetes dataset contains 8 different attributes for dataset description. On the basis of these attributes prediction of diabetes to a person has been analyzed. For security purpose arithmetic and trigonometric formulas have been implemented on values of all the attributes so that originality of data gets changed. After encryption of the dataset encrypted data has been used for classification based on rule based and tree based classification algorithms, these algorithm generated rules and splitting trees from the training dataset and used these to divide testing dataset into different classes.

In the purposed work various parameter that are accuracy, precision, recall, F-measure, ROC, TP rate and TN rate has been measured. On the basis of these parameters one can conclude that dataset using arithmetic and trig metric formulas provides better classification results as compare to original data.

References

[1] Manjari Anand ―Customer Relationship Management using Adaptive Resonance Theory", International Journal of Computer Applications, 2013, pp. 43-47.
[2] Ms. Saranya, **"**Decision Support System for CRM in Online Shopping System**"**, International Journal of Advances in Computer Science and Technology, 3(2), February 2014, 148, 2014, pp. 148-151.
[3] Kamal R. ―Adaptive Pointing Theory (APT) Artificial Neural Network", International Journal of Computer and Communication Engineering, 2014, pp. 212-215.
[4] Mohammed Al-Maolegi ―An Improved Apriori Algorithm For Association Rules", International Journal on Natural Language Computing (IJNLC), 2014, pp. 21-29.
[5] Meenakshi ―Survey on Classification Methods using WEKA", International Journal of Computer Applications, 2014, pp. 16-19.
[6] Wei Wang., ―Application of Data Mining Technique in Customer Segmentation of Shipping Enterprises", IEEE International Conference on Database Technology and Applications, 2010, pp. 1-4.
[7] E. W. T. Nagy ―Application of data mining techniques in customer relationship management: A literature review and classification" IEEE Conf. on Expert Systems with Applications 2009, pp. 2592–2602.
[8] Asghar, S. ―Automated Data Mining Techniques: A Critical Literature Review" 978-0-7695-3595-1, pp. 75–79, IEEE, 2009.
[9] D. Clot., ―Using functional PCA for cardiac motion exploration", IEEE International Conference on Data Mining, 2002, pp. 91-98.

[10] Yin-Fu Huang., —Mining generalized association rules using pruning techniques", IEEE International Conference on Data Mining, 2002, pp. 227–234.
[11] Yuh-Jyh Hu., —Mining a set of co regulated RNA sequences", IEEE International Conference on Data Mining, 2002, pp. 625-628.
[12] P. Cohen., —Unsupervised segmentation of categorical time series into episodes", IEEE International Conference on Data Mining, 2002, pp. 99-106.