A Novel Scheme for Securing Medical Data by using Hybrid Privacy Preserving Mechanism in Healthcare Application

M.Rameshkumar¹, Dr. V.Lakshmi Prabha²

¹Ass. Professor, Department of Computer Science, Govt. Arts College, Nandanam, Chennai, TN, India
²Asst. Professor, Department of Computer Science, Rani Anna Govt. Arts College, Tirunelveli TN, India

Abstract—Data mining techniques help clinician make proper decisions in health care applications. The benefits of clinical decision support system include not only improving diagnosis of illness accuracy but also reducing identification time. Specifically, with massive amounts of clinical data generated every day, decision tree algorithm can be utilized to excavate valuable data to improve clinical decision support system. In this paper a hybrid algorithm is proposed which integrates both the homomorphic encryption and naive bayes for storing the data in database safe manner. In order to retrieve the data safely and classify the abnormal and normal data fuzzy based system is employed here.

Keywords - encryption, Privacy preserving, technique, Navie Bayesian, Fuzzy logic.

I. INTRODUCTION

Data mining is that the methodology of analysing info from utterly completely different views and summarizing it in to useful information, data processing is wide used in varied applications like business (insurance, banking, retail), science analysis (astronomy, medicine), and certificate (detection of criminals and terrorists) immense amount of knowledge area unit store at intervals the knowledge and sadly, these giant collections of knowledge keep it up distinct structures really chop-chop became overwhelming.

This initial chaos has crystal rectifier to the creation of structured databases and management systems (DBMS)[1]. The economic management systems square measure necessary assets for management of associate degree oversized corpus associate degree for effective and economical retrieval of specific info from an oversized assortment whenever needed. The data discovery info is associate unvarying methodology. The key reason that mining has attracted giant measure of attention at intervals business at intervals the recent years is owing to the wide accessibility of big amounts and additionally they would love for turning such data into useful information and data. The data gained are going to be used for applications ranging from business management, production management, and market analysis to rising vogue and science exploration and health info analysis [2]. Information has become associate organization’s most precious quality. Organizations became a lot of and a lot of obsessed with information, since more information is being hold on and processed on network-based systems. a significant challenge in providing associate a good and economical protective mechanism to a network is that the flexibility to note novel attacks or any intrusion works and implement counter measures. Intrusion detection could also be a significant part in securing information systems whereas it's is implemented by associate Intrusion detection system. Today, we tend to are ready to notice many industrial Intrusion Detection Systems gettable inside the market, but they are restricted in their observation utility which they require frequent updates and patches. The wide unfold use of ecommerce, has raised the necessity of protecting the system to a very high extend [3]. Confidentiality, Integrity and accessibility of knowledge are major concerns inside the event and exploitation of network primarily based computer systems. Intrusion Detection System, can detect, forestall and react to the attacks. Intrusion Detection has become associate integral a district of the info security methodology. However, it isn't technically potential to form a system with no vulnerabilities; intrusion detection continues to be a significant area of study.

Host-based systems use native host data like method behavior; file integrity and system logs to observe events. Network-based systems use network activity to perform the analysis. combos of those two varieties are potential. reckoning on however the intrusion is detected associate IDS is any classified as signature-based (also referred to as misuse system) or anomaly-based [7]. Signature-based systems arrange to match determined activities against well outlined patterns that additionally known as signatures. Anomaly-based systems explore for any proof of activities that deviate from what's thought of traditional system use. These systems are capable of sleuthing attacks that a well-defined pattern doesn't exist. A hybrid IDS is capable of exploitation signatures and sleuthing anomalies [4]. a lot of
recently, techniques from knowledge of mining space (mining of association rules and frequency episodes) are accustomed mine the conventional patterns from audit data. Typically, associate IDS uses mathematical logic in determinant whether or not an associate intrusion is detected and also the use of formal logic has been investigated as an alternate to mathematical logic within the style and implementation of those systems. Formal logic addresses the formal principles of approximate reasoning [5]. It provides a sound foundation to handle inexactness and uncleanness likewise as mature logical thinking mechanisms exploitation variable degrees of truth. Since boundaries don't seem to be continually clearly outlined, formal logic will be accustomed determine complicated pattern or behaviour variations [6]. This can be accomplished by building associate Intrusion Detection System that mixes formal logic rules with associate knowledgeable system responsible of evaluating rule certainties.

The evolution of data mining technique has obligatory a serious impact on the revolution of human’s fashion by predicting behaviours and future trends on everything which might convert keep data into purposeful information. Artificial intelligence (AI) techniques are applied to each misuse detection and anomaly detection. Rule primarily based skilled systems have served because the basis for many systems as well as SRI’s Intrusion Detection skilled System (IDES)[2]. These systems encrypt associate degree expert’s information of known patterns of attack and system vulnerabilities as if-then rules. The acquisition of those rules may be a tedious and fallible process; this downside (known because the information acquisition bottleneck in skilled system literature) has generated a good deal of interest within the application of machine learning techniques to alter the method of learning the patterns.

Examples symbolize the Time-based Inductive Machine (TIM) for intrusion detection [3] that learns successive patterns and neural network-based intrusion detection systems [4]. Additional recently, techniques from knowledge mining space (mining of association rules and frequency episodes) are accustomed mine traditional patterns from audit data [5, 10, 15]. Issues are encountered, however, if one derives rules that are directly keen about audit information [6] associate degree intrusion that deviates solely slightly from a pattern derived from the audit information might not be detected or little modification in traditional behavior might cause a warning.

II. LITERATURE REVIEW

The IDS approaches are typically categorised into misuse and anomaly detection approaches within the literature. Researchers within the past have investigated regarding this subject. Yihua Liao et al., (2002) projected associate degree approach supported the k-Nearest Neighbor (kNN) classifier that is employed to classify the program behavior as traditional or intrusive. Within the kNN classifier, the frequencies of system calls are wont to describe the program behavior. Also, text categorization techniques were adopted to convert every method to a vector and calculate the similarity between two program activities [12]. Hence, there's no have to be compelled to learn individual program profiles one by one then a lot of calculations are primarily reduced.

Jian gallinacean et al [13] projected formula that reduced the warning rate IDS activities with symbolic logic and data processing formula. a collection of fuzzy rules is wont to classify the traditional and abnormal activities during a network, associate degree fuzzy data processing algorithms is applied over fuzzy rules to work out once an
intrusion is current. The mine fuzzy association rules and fuzzy frequent episodes are explained for anomaly detection from audit knowledge and therefore the experimental results indicate that fuzzy data processing will give effective approximate anomaly detection.

Xiao et al (2005) [14] bestowed a method for intrusion discovering that applies GA to detect intrusion in networks through effective feature choice. Their approach uses scientific theory to extract relevant options and scale back the complexity. Then, they fashioned a linear structure rule from the chosen options so as to classify network behaviors into traditional and abnormal behavior. However, their approach considers solely distinct options.

The Least Squares Support Vector Machines (LSSVM) model was advised by Haihua federal agency et al (2006), mistreatment kernel house approximation through greedy looking out and therefore created a topological space basis of original space inhabited by coaching knowledge. By means that of this approximation, the coaching knowledge was downsized and consequently, the information of support vectors of LSSVM model were reduced [15]. So, the effect time of intrusion detection was enhanced. The model has been evaluated mistreatment KDD Cup99 knowledge and therefore the results demonstrate that the tactic is a good approach for quick intrusion detection.

Han-Ching Chinese et al (2010) projected associate degree approach that produces use of the analytical strengths of neural networks to discover stepping-stone intrusion with two schemes. Along with the two schemes, one uses eight packet variables and therefore the alternative clusters a sequence of consecutive packet in round-trip times.[16] varied transfer performs and learning rules were analyzed and therefore the authors all over that mistreatment sigmoid transfer function and delta learning rule usually provides higher prediction.

Hu et al (2009) projected an easy knowledge pre-processing approach to hurry up Hidden Andrei Markov Model coaching for system-call-based anomaly intrusion detection. Coaching and attack knowledge are from a public info provided by the University of Recent North American Nation and therefore the Massachusetts Institute of Technology computer science Laboratory.

Maggi et al (2010) projected associate degree unsupervised host-based intrusion detection system supported supervisor call instruction arguments and sequences. They outlined a collection of anomaly detection models for the individual parameters of the decision so describe an agglomeration method that helps to higher work models to supervisor call instruction arguments and creates interrelations among completely different arguments of a supervisor call instruction. Finally, they used the behavioural Andrei Markov model so as to capture time correlations and abnormal behaviors. The complete system desires no previous information input.

John Felix Charles Joseph et al (2010) projected a technique, specifically Cross layer primarily based adaptation period routing Attack Detection System for adapting the intrusion detection model at real time with two major stages. Within the initial stage, the incidence of recent patterns within the routing management traffic was known and prioritized supported their data content. The second stage of adaptation was to incrementally update the detection model mistreatment the new patterns with minimum process overhead.

Nong Ye et al (2013) classified the attacks during a comprehensive and smart format and explained the System-Fault-Risk outline for cyber attack classification. The bottom of the frame work is on a scientific foundation, combining theory from classification engineering, fault modeling, and risk-assessment. They advised that a company ought to perceive the attacks by assessing them mistreatment the metrics like attack characteristics, activities, state and performance impact that helped in selecting effective barriers and may decide the preparation position of such barriers. Privacy-preserving parallel distributed classification techniques [18] wherever multiple sites collaborate and broadcast the mining results. But within the method, no data regarding either the knowledge maintained within the sites or data obtained throughout computation is divulged. Two protocols are bestowed to construct a Privacy protective Naïve theorem classifier mistreatment the Pailler’s homomorphic encoding techniques. The issues of association rule mining wherever transactions distributed across sources are explained in [17] every web site holds some attributes of every group action, and therefore the sites want to collaborate to spot globally valid association rules.

III.RESEARCH METHODOLOGY

The main objective of the proposed work is to secure the medical data from the unauthorized users and classify the concern data in order to reduce time and to achieve accuracy. In order to retrieve data fuzzy set rules are followed to classify the normal and abnormal data. Figure 2 demonstrates of the general architecture for the proposed work. The proposed framework is sub divided into three stages: authentication, clustering and fuzzy based classification.

The primary stage executes authentication and authorization, whereas the authorized user have separate key to access the data in order to do that an efficient method of hybrid data mining technique using here which includes the integration of homomorphic encryption algorithm and navie bayesian classifier. Hybrid data mining technique plays a vital role for storing the data in database, in the case of retrieving data safely from database.
A. Filterative association rule

It uses a breadth-first search strategy to investigating the support of itemsets and uses a candidate generation perform that exploits the downward closure property of support. It's a collection of concepts supported the frequency of two stage approach which may be rotten into two sub-problems 1st, notice all Frequent Itemsets that have dealing supports on top of minimum support known as minsup. The support for associate itemset is printed as a result of the fraction of total transactions that contains this itemset. Itemsets with minimum support are referred to as large itemsets, and each one the others very little itemsets. Second, use the big itemsets to come up with the specified rules with the algorithmic technique.

Steps in association rule

Step1: scan the data base (D) for count of each candidate and generate the list C(i). List all the candidate item sets and place the corresponding candidate support count in L(i).

Step2: compare the candidate support count with the minimum support count. Generate candidate C(i) from L(i).

Step3: scan D for count of each candidate. Compare the candidate support count with the minimum support count list remaining in L(i) .This process is continued until the most frequent item set is produced.

B. Homomorphic algorithm

Let (A, B, C, D) be a encryption scheme (where A is the function to generate public parameters, B and C are the encryption and decryption functions, and D is the message space respectively) with the following properties:

- The encryption scheme (A, B, C) is semantically secure [24]. Essentially, an encryption scheme is semantically secure if an adversary gains no extra information by inspecting the ciphertext.
- For all g ∈ G and α ∈ G, gα1 ∈ G(α1). Encrypting the same message twice in a probabilistic encryption function can yield a different ciphertext, so G(m) denotes the set of cipher texts that can be obtained by encrypting m1.
- There is a computable function f such that for all messages d1 and d2 the following property holds: f(E(m1), E(m2)) = E(m1 + m2).

Steps for Homomorphic Encryption

- (Step 1) Party 1 encrypts x and n and sends the encrypted values y1 ∈ B(y) and n1 ∈ B(n) to party 2.
- (Step 2) Party 2 computes a random message z ∈ B, and encrypts y(z · x) and y(n · z) to obtain z1 ∈ B(y(z · x)) and z2 ∈ B(z · n). Party 2 computes the following two messages and sends it to party 1:
  
  d1 = f(y z 1 , z1)
  d2 = f(n z 1 , z2)

Note: In our implementation we use the homomorphic-encryption scheme by [5] where f is multiplication.

- (Step 3) Using the two properties of the probabilistic encryption scheme (A,B,C), we have the following:
  
  d1 = B(z · x + z · y)
  d2 = B(z · n + z · m)

Therefore, party 1 can compute y(x + y) and y(n + m), and hence can compute x+y n+m. Party 1 sends x+y n+m to party 2.

C. Naive Bayesian

The Naive Bayes (NB) formula makes predictions victimization Bayes Theorem that derives the likelihood of a prediction from the underlying proof. NB affords quick model build. NB assumes that every attribute, or piece of proof, is

---

User request

Filterative association rule

Data base

Hybrid data mining technique

Fuzzy set rules

Abnormal data

Normal data

Navie bayesian classifier

Authentication and authorization

Performance analysis

Figure 2. System architecture of the proposed system

fuzzy set of rules will be followed to classify the normal and abnormal data.
freelance from the others. In follow, this assumption sometimes doesn't degrade the model's prophetic accuracy considerably, it's a probabilistic graphical model that represents a collection of variables and their probabilistic independencies.

Bayes theorem provides how of scheming the posterior chance, \( A(b|y) \), from \( A(c) \), \( A(x) \), and \( A(y|b) \). Naive mathematician categoryifier assumes that the result of the worth of a predictor (x) on a given class (c) is freelance of the values of different predictors. This assumption is termed category conditional independence.

\[
\begin{align*}
  A(B \mid Y) &= A(Y \mid B) \cdot A(B) \\
  A(b|y) &= \text{the posterior chance of sophistication (target) given predictor (attribute)}.
\end{align*}
\]

A(b) is that the previous chance of sophistication.

A(y|b) is that the probability that is that the chance of predictor given category.

A(y) is that the previous chance of predictor.

In this rule initially we discover out the previous chance for the given information set then conclude category probability for the info set. Subsequently we discover out the very best classifier chance and base on we discover detection rate and false positive for the intrusion information set. The previous chance A(Bm) for every category is calculable by tally however typically each category happens within the dataset Di. For every attribute Ai the amount of occurrences of every attribute price Amn are often confirm A(Am). The probability \( P(\text{Amn}/\text{Ch} ) \) for every attribute prices Amn may be calculable by tally however often every attribute value happens within the class within the dataset D.

**Steps to detect intrusion in Navie bayesian**

**Step 1:** Take the Class CL From D.
**Step 2:** For each attribute value Remove the noise from the dataset.
**Step 3:** Calculate the prior probability \( P(Bi) \) for each class Bi in dataset.

\[
D: P(Bi) = \sum_{i=1}^{n} t_i \rightarrow bi
\]

**Step 4:** Calculate the class conditional probabilities \( P(Aij| Bi) \) for each attribute values in dataset D
**Step 5:** Multiply the prior probability and class conditional probability.
**Step 6:** Consider the class with the highest classifier probability.
**Step 7:** Repeat steps 2 to 4 until all attribute at their highest probability.

**D. Fuzzy based classification**

By exploitation the fuzzy set of rules traditional knowledge and abnormal knowledge are often classified whereas threshold worth is about to perform the classification. The method of a fuzzy system has 3 steps. These steps square measure Fuzzification, Rule analysis, and Defuzzification. Within the fuzzification step, the input crisp values square measure reworked into degrees of membership within the fuzzy sets. Within the rule analysis step, every fuzzy rule is assigned with a strength worth. The strength is decided by the degrees of memberships of the crisp input values within the fuzzy sets of antecedent a part of the fuzzy rule. The defuzzification stage transposes the fuzzy outputs into crisp values. The intrusion fuzzy set and non-intrusion fuzzy set is employed to come up with fuzzy rules exploitation the Fuzzy Rule. This rule formulates fuzzy rule exploitation the information gained through analysis of instances that helps it to discriminate the conventional and abnormal knowledge. The fuzzy rule places the complete traditional and abnormal knowledge in a very separate set or a vector by exploitation the pseudo code. And therefore once knowledge is received the symbolic logic itself will classify the regular data from the deviant one.

**Pseudo code generation: Fuzzy set algorithm**

**Step 1:** Selected the attribute (A) for identifying Intrusion/non intrusion cluster
**Step 2:** Subset of operation (S) in order to store unique values
**Step 3:** Selects the next element from the available data (k)

\[
S= \text{item}[i]
\]

For \( i=1 \) to \( n-1 \)

\[
K= \text{DataField}[i+1]
\]

\[
S= S \cup K
\]

Select unique item of the field
**End for**

**Step 4:** Store S

**IV. PERFORMANCE ANALYSIS**

To evaluate the performance of the planned model the sensitivity, specificity, and accuracy are calculated. The sensitivity is that the proportion of positive instances that are properly classified as positive (e.g. the proportion of sick folks that are classified as sick). The specificity is that the proportion of negative instances that are properly classified as negative (e.g. the proportion of healthy folks that are classified as healthy). The accuracy is that the proportion of instances that are properly classified.

\[
\text{Sensitivity} = \frac{\text{True Positive}}{\text{Positive}} \quad \text{Specificity} = \frac{\text{True Negative}}{\text{Negative}} \quad \text{Accuracy} (\%) = \frac{(\text{TP} + \text{TN})}{(\text{Positive} + \text{Negative})}
\]
V. CONCLUSIONS

Disease prediction is the leading cause of death all over the world. Researchers have been investigating applying various data mining techniques to help health care professional in the diagnosis of patient’s disease. Fuzzy logic and naieve bayesian are the successful data mining techniques used in the diagnosis of patient’s disease. This paper investigated integrating Bayesian classifier with Homomorphic algorithm will provide the security for validating the patient data and decision tree for analysing the patient disease with more accuracy. The results show that integrating naive Bayesian classifier with Homomorphic algorithm and fuzzy logic can accuracy and computational time in the diagnosis of patient’s disease.
REFERENCES


