IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Deriving Attribute Relationship for Big Data Security

Bindiya M.K[1], Dr. RaviKumar G.K[2]

Research Scholar, SJBIT, Benagaluru[1]

Wipro Technologies, Bengaluru[2]

Abstract: There has been an ever rising interest towards Big Data and its security. Big Data is vast to be analysed and providing security to entire Big Data is an uphill task. Finding a solution to make sure that the Big Data is secured could be easier when we try to select only important information from the Big Data and provide security to that information only. Hence, the main goal of this paper is to propose an attribute selection methodology based on the relationship and the sensitivity among attributes for Big Data security. To be more precise, this paper basically includes two things: attribute selection methodology and provision of security. When it comes to providing security, one must understand that not all the data in Big Data needs to be secured because not every information in the Big Data is rather important. So this paper tries to select an attribute based on its relevance and the confidentiality level and provide security to that particular attribute only. That is, an attribute with higher relevance is more important than other attributes to provide security. This process includes two Datasets: User-defined Dataset and Reference Dataset. Attributes from both the Datasets are compared for their relevance and these attributes are classified into three categories namely Restricted, Confidential and Public. This becomes easier to provide security to each of these categories accordingly.

Keywords: Big Data, Machine Learning, Artificial Neural Network, Data Mining, Datasets, Security

I. INTRODUCTION

In today's technology, the cloud and big data are major advanced. A technique of big data is to process an drive a small amount of data from large amount of data in an existing database. Here the value is extracted in a variety of ways. Big data can be defined as it is an combination of structured , unstructured and semi structured of data. It has an ability to extract useful information. Big data can be defined by 4 V's,means velocity,volume,variety and validity. Volume defines by the size of data, Velocity define the speed at which the data can be processed, and variety define the types of data and validity define how long the data is valid. Although,the quantity of big data can't be speak about petabytes to Exabyte of data and it cannot be easily integrated .For loading big data into traditional relational database, it consumes more time and also too much cost, so it require new approaches to analyzing and loading the data that must rely on less quality of data and data chema.Instead,using Artificial intelligence(AI) ,data lake and machine learning concepts that use complex algorithms to check for repeatable pattern from raw data.

Hadoop is a platform that used to process and store large volume of data sets that are distributed across a multiple clusters and map reduce is a programming technique in which it is used to combine and process data which are distributed across multiple sources. Cloud computing associated with big data analytics requires this platform to analyze large volume of data sets in real time. An axiom of big data is that "small data is for people, Big data is for machines". Big data can often used to describe data whose size of data and format can be used for analytics in self-service manner.The rest of the paper is organized as follows: Section II, gives introduction and the background study of data mining, Section III discusses the concept of artificial neural network. The system architecture for the training and testing phase along with the resultsis presented and discussed in Section IV. Lastly, conclusion and future work discussed in Section V.

II. BACKGROUND

Examining large amount of data from database in order to produce a useful information is known as Data minig.In other words we can also define that minig the knowledge of data from huge sets of data.Today,in information industry large amount of data is available and this data has to be converted into useful information. In order to convert this raw data into useful information, it requires analytics to extract information.

DOI10.17148/IJARCCE.2017.61046

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction. Interesting measures and thresholds for pattern evaluationis used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge. Representation for visualizing the discovered patterns, refers to the form in which discovered patterns are to be displayed. They may be rules, tables, charts, graphs, etc. Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. Techniques to discover the hidden patterns in large data sets and focusing on issues related to feasibility, usefulness, effectiveness and scalability [8]. Use of Non RDBMS for storing and retrieving the data [2]. Limited access to Big Data; Not all data are equivalent [6]. Data mining techniques are used for data in rest and data in motion [5]. Some of the drawbacks of the existing system are: (a) Higher overhead cost and greater computational complexity. (b) Old and obsolete hardware, which has less processing speed and the response is slow. (c) Also the Naïve Bayes inference method of clustering is used which is less efficient.

III. ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a computational model based on the structure and functions of biological neural network i.e. the human brain. It's a mesh like network that is made up of millions of neurons. The brain basically learns from experience. Similarly the information that flows through the network affects the structure of the ANN because a neural network changes or learns based on the input and output.ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled or patterns are found. ANN is also known as a neuron network. An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function approximation tool. These types of tools help to estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distributions. ANN takes data samples rather than entire data sets to arrive at solutions, which saves both time and money. ANNs have three layers that are interconnected.

The first layer consists of input neurons. Those neurons send data on to the second layer which is usually called the hidden layer. This is the layer where activation functions are present. They are responsible to learn the input, apply the function and then produce the output which in turn is sent to the output neurons of the third layer. Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms and activation functions.

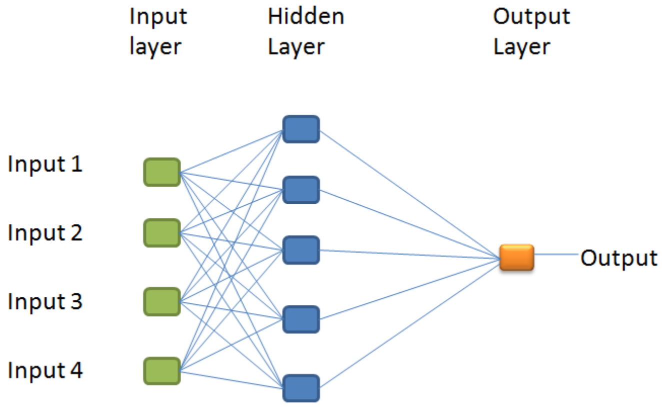


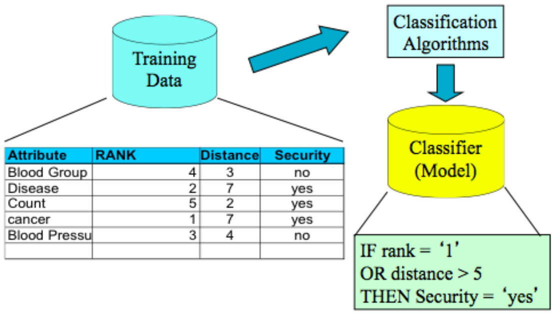
Figure 1 - Artificial Neural Network

IV. SYSTEM ARCHITECTURE

A. Training Phase

Consider the two Datasets, one which is the Reference Dataset and the other which is collected from various sources i.e. the User- defined Dataset. Here, medical datais being taken into account, in specifica simple blood analysis as this needs to be protected with the highest level of security. A simple blood analysis can reveal the entire genetics of a person and it contains some information which is not meant to be shared with anybody. For example: HIV, cancerous properties, risk of a cardiac problem.

DOI10.17148/IJARCCE.2017.61046



Attribute	RANK	Distance	Security
Blood Group	4	3	no
Disease	2	7	yes
Count	5	2	yes
cancer	1	7	yes
Blood Pressu	3	4	no

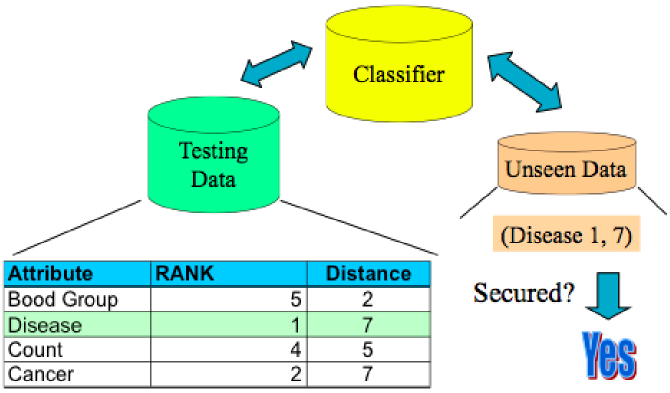
Figure 2 - Training Phase

The training data in Figure 2 consists of attributes related to the blood like blood group, blood count, blood pressure, cancerous properties and diseases. Here rank 1 has the highest priority and rank 5 has the lowest priority. The distance is calculated based on how relevant and similar the attributes are with respective to each other in the two Datasets.

Example: Consider the blood group attribute. In scenario 1: if the blood group is B+ve in the user-defined dataset and also B+ve in the reference dataset, we call this relation equivalent and assign it a distance of 5. In scenario 2: if the blood group is B-ve in the user-defined dataset and B+ve in the reference dataset, we call this relation hierarchical and assign it a distance of 3. In scenario 3: if the blood group is A in the user-defined dataset and B+ve in the reference dataset, we call this relation non-equivalent and assign it a distance of 1. This is how the distance matrix is created. The distance matrix is fed into the classifier. The maximum value of the entire row is the basis on which the data is classified into Restricted, Confidential and Public. For n number of data, we run the iteration n number of times just to improve the accuracy. We use BPA or Back Propagation Algorithm for training the network

B. Testing phase

Here in Figure 3, we are providing the machine which has already been trained with the input i.e. the test data. It contains only the attribute, rank and distance fields. The machine will now correctly and accurately predict the outputi.e. the data classified into Restricted, Confidential and Public. Only the restricted data is secured using the encryption and decryption technique.



Attribute	RANK	Distance
Bood Group	5	2
Disease	1	7
Count	4	5
Cancer	2	7

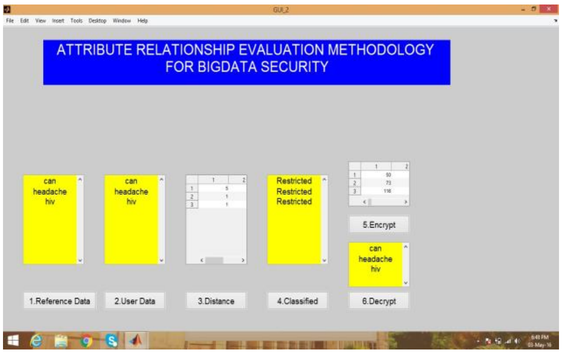
Figure 3- Testing Phase

C. Security

Since 'Restricted' category is the one which requires higher security, we use simple and widely known Security technique called RSA algorithm to provide security to that category. But to increase the security level, instead of creating single-part cipher text, we create multi-part cipher text. As in, plain text at one point will be converted into cipher text and is distributed into 128 parts. During decrypting the cipher text, one needs to gather all 128 parts to form one cipher text and can use his private key to decipher the same.

D. Results

Test Case 1:



ATTRIBUTE RELATIONSHIP EVALUATION METHODOLOGY FOR BIGDATA SECURITY

can
headache
hiv

can
headache
hiv

	1	2
1	5	
2	1	
3	1	

Restricted
Restricted
Restricted

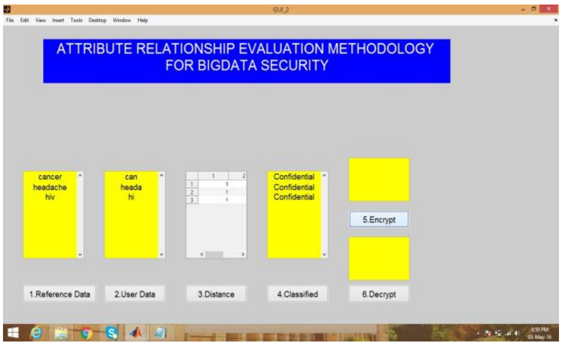
	1	2
1	50	
2	73	
3	116	

can
headache
hiv

Figure 4 – Equivalent

In Figure 4, we consider both the reference data and the user data to contain the same data. The distance is calculated based on the relevance between the reference data and the user data. Since we have same data, we call this relation equivalent. Then the artificial neural network is trained and it classifies the data as 'restricted' in this case. As per our proposed system, we include security only to restricted data and that is what is happening in the above snapshot. The required user can then decrypt the restricted data.

Test Case 2:



ATTRIBUTE RELATIONSHIP EVALUATION METHODOLOGY FOR BIGDATA SECURITY

cancer
headache
hiv

can
heada
hi

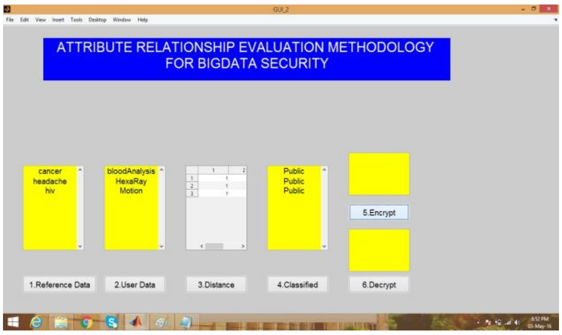
	1	2
1	3	
2	1	
3	1	

Confidential
Confidential
Confidential

Figure 5 – Hierarchical

In Figure 5, we consider the reference data to contain a part of the user data. The distance is calculated based on the relevance between the reference data and the user data. Since a portion of the user data matches the reference data, we call this relation hierarchical. Then the artificial neural network is trained and it classifies the data as 'confidential' in this case. As per our proposed system, we include security only to restricted data and therefore in the above snapshot no security is being given.

Test Case 3:



ATTRIBUTE RELATIONSHIP EVALUATION METHODOLOGY FOR BIGDATA SECURITY

cancer
headache
hiv

bloodAnalysis
HexaRay
Motion

	1	2
1	1	
2	1	
3	1	

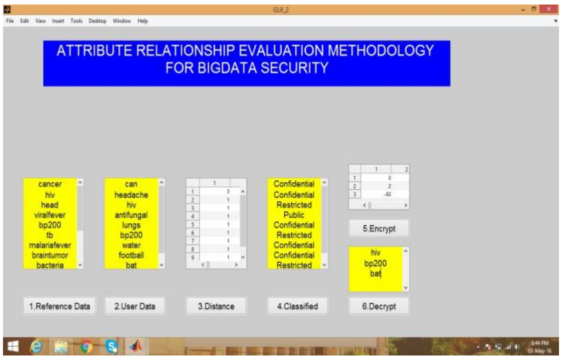
Public
Public
Public

Figure 6 – Non-Equivalent

DOI10.17148/IJARCCE.2017.61046

In Figure 6, we consider the reference data and user data to contain different data i.e. nothing common. The distance is calculated based on the relevance between the reference data and the user data. Since nothing is similar between both the user data and the reference data, we call this relation non-equivalent. Then the artificial neural network is trained and it classifies the data as 'public' in this case. As per our proposed system, we include security only to restricted data and therefore in the above snapshot no security is being given.

Test Case 4:



ATTRIBUTE RELATIONSHIP EVALUATION METHODOLOGY
FOR BIGDATA SECURITY

cancer
hiv
head
viralfever
bp200
tb
malariafever
braintumor
bacteria

can
headache
hiv
antifungal
lungs
bp200
water
football
bat

	1
1	3
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1

Confidential
Confidential
Restricted
Public
Confidential
Restricted
Confidential
Confidential
Restricted

	1	2
1		2
2		2
3		-82

hiv
bp200
bat

Figure 7 – Mixed

In Figure 7, we consider the reference data and user data to have a combination of the above mentioned cases. The distance is calculated based on the relevance between the reference data and the user data. Then the artificial neural network is trained and it classifies the data into restricted, confidential and public based on the similarity between the user data and the reference data. As per our proposed system, we include security only to restricted data and that is what is being done in the above snapshot.

V. CONCLUSION AND FUTURE WORK

In this paper, weused the concept of Artificial Neural Network to train the machine and select the attributesbased on the different levels of security i.e. Restricted, Confidential and Public from the User- defined dataset and the Reference dataset. Security was imposed on the data which showed the highest level of sensitivity i.e. 'restricted' in our case. Results were obtained at a faster rate and with the highest level of accuracy possible.

In future, one can plan to introduce the Support Vector Machine algorithm, which is more advantageous because of its speedy prediction time.Also different layers of security can be included for Restricted, Confidential and Public data. This would make the system less prone to threats and attacks. Since this paper only considers text data, one can also extend this technique to image, audio and video data as well.

REFERENCES

[1] M. Chen, J. Han and P.S. Yu, "Data mining: An overview from a database perspective," Knowledge and data Engineering, IEEE Transactions, vol. 8, no. 6, pp. 866-883 1996.
[2] C. Strauch, U.S. Sites and W. Kriha, "NoSQL databases," URL: http://www.christof-strauch.de/nosqldbs. 2011.
[3] J. Bughin, M. Chui and J. Manyika, "Clouds, Big Data, and smart assets: Ten tech-enabled business trends to watch," McKinsey Quarterly, vol. 56, 2010.
[4] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big Data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, pp. 1-137, 2011.
[5] R. Gupta, H. Gupta and M. Mohania, "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?" in Big Data Analytics, Anonymous: Springer, 2012, pp. 42-61. , 2012.
[6] D. Boyd and K. Crawford, "Six provocations for Big Data," 2011.
[7] F. Gorunescu, Data Mining: Concepts, models and techniques, Springer, 2011.
[8] J. Han, M. Kamber and J. Pei, Data mining: concepts and techniques, Morgan Kaufmann, 2006.
[9] S. H. Kim, N. U. Kim, T. M. Chung,"Attribute Relationship Evaluation Methodology for Big Data Security", IEEE, 2010, pp. 1-4

DOI10.17148/IJARCCE.2017.61046