ABSTRACT:

The world has been showing an ever rising interest towards Big Data and Big Data security. But one should understand that the Big Data is just an extension of Data mining, but not a new technology itself. However, the main goal of this paper is to propose an attribute selection methodology for Big Data security. It also includes two things: attribute selection and providing security. Providing security, means one must understand that not all the data in Big Data needs to be secured because not every information in the Big Data is 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 which has higher relevance is more important than other attributes to provide security. This paper basically considers two Datasets: User-defined Dataset and Reference Dataset. Attributes from both the Datasets are compared for their relevance and classify these attributes into three categories namely Restricted, Confidential and Public. This becomes easier for the user to provide security to each of these categories accordingly.

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

[1] INTRODUCTION

Big Data and Cloud are a major trend in today’s technologically advanced world. Big Data is a technique to extract a small value from the huge amount of enormous data beyond
the processing capabilities of existing database. Here the value is extracted in a variety of ways.

Because of the sheer amount and complexity of the information available, for example from, weather data, geographical data, drug testing data, health care imaging information systems, atomic particle accelerator detector data, world wide web pages, e-commerce data, etc., nowadays, engineers and scientists rely heavily on computers to process and analyse data. This is why Machine Learning (ML) has become an emerging topic of research that has been employed by an increasing number of disciplines to automate complex decision-making and problem-solving tasks. This is because the goal of Machine Learning (ML) is to extract knowledge from experimental data and to use computers for complex decision-making. Machine learning techniques can be divided into three major groups based on the types of problems they can solve, namely, supervised, semi-supervised and unsupervised learning. Intuitively, it is not possible to protect all the data inside the Big Data. The objective of this paper is to select the attributes that should be protected in the Big Data. For that we need to apply the Machine Learning algorithm to develop a system which secures a particular attribute which is of more sensitivity as far as the security is concerned. The Artificial Neural Network (ANN) is trained with the parameters like the rank and the distance of the inter Big Data attribute relevance measure.

The rest of the paper is organized as follows: Section II, gives introduction and the background of Big Data, Section III discusses the concept of artificial neural network. The proposed system architecture for the training and testing phase along with the creation of relevance matrix is presented and discussed in Section IV. Lastly, conclusion and future work discussed in Section V.

[2] BACKGROUND

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

The background knowledge allows data to be mined at multiple levels of abstraction. 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 Naive Bayes inference method of clustering is used which is less efficient.
[3] ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes or learns based on that 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.

[4] PROPOSED SYSTEM ARCHITECTURE

The contributions in this paper are:

- The Artificial Neural Network just like the Machine learning algorithm is robust and it is used to predict whether an attribute needs to be secured or not.
- Replacing the hard threshold techniques, which would lose the proper orthogonality in the data while the decision-making is applied, the Artificial Neural Network is used where there is more scope of orthogonality.
- The ANN based training and testing is applied to get a very robust relevance-mapping algorithm for application of the security measures on the attributes.

[4.1] 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 data is being taken into account, in specific a 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.

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 cancer attribute in this example takes the highest priority hence it is assigned to rank 1. Similarly the disease takes the next highest priority which is assigned to rank 2 and so on. The distance is calculated based on how relevant and similar the attributes are with respective to each other in the two Datasets.

In the classifier, the machine is trained in such a way that if either the rank is 1 or the distance is greater than 5, the attribute needs to be secured. For n number of data, we run the iteration n number of times just to improve the accuracy. Now just considering the rank field, the classifier will segregate the data based on the different levels of security.

**[4.2] 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 output i.e. the attribute which needs to be secured. It will also classify the data based on different levels of security which is restricted, confidential and public. Now this data is available to add the necessary security features.

**[4.3] Creation of Relevance Matrix**

Representation of Big Data with attributes centrally for using relationship between attributes is as follows:

\[ DS_i = \text{Any object type Big Data } 'i' \]

\[ \text{Attr}_{p} = 'p' \text{ Attribute of Big Data} \]
\[ DS_i(Attr) = \text{Attribute set of Big Data \textquoteleft DS}_i \]
\[ DS_i[Attr_p] = \text{Length of Attribute \textquoteleft p\textquoteleft of Big Data \textquoteleft i\textquoteleft} \]

For instance, there are two Datasets. It can be represented as shown in the following table.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>DS 1</td>
<td>Attr_p</td>
</tr>
<tr>
<td>DS 2</td>
<td>Attr_m</td>
</tr>
</tbody>
</table>

Table: 1. Example of Big Data

Suppose that all attributes have some relationship with each other. We classified attribute correlation into three cases.

1. **Equivalence Relation**: Two attributes are the same.
   \[ Attr_p \equiv Attr_q \]

2. **Hierarchical Relation**: Two attributes have hierarchical relationship.
   \[ Attr_p \not\equiv Attr_q \land Attr_p \subset Attr_q \lor Attr_p \supset Attr_q \]

3. **Unknown Relation**: Neither Equivalence nor Hierarchical.
   \[ Attr_p \not\equiv Attr_q \land Attr_p \not\subset Attr_q \land Attr_p \not\supset Attr_q \]

At this time, ‘equivalence’ relation is given the highest value and ‘unknown’ is given the lowest value. First, find all attributes of object type Big Data. And then compare the attribute relationship between mutual Datasets.

Let the two Datasets be \( D_1 \) and \( D_2 \)
\[ DSD1(Attr) = \{Attr_m, Attr_n, Attr_o, Attr_t\} \rightarrow 4 \text{ attributes} \]
\[ DSD2(Attr) = \{Attr_m', Attr_n, Attr_j\} \rightarrow 3 \text{ attributes} \]

Next step is comparing the attribute relationship between mutual Datasets using the table.

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attr_m</td>
</tr>
<tr>
<td>D2</td>
<td></td>
</tr>
<tr>
<td>Attr_m</td>
<td>Hierarchy</td>
</tr>
<tr>
<td>Attr_n</td>
<td>Unknown</td>
</tr>
<tr>
<td>Attr_j</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Table: 2. Comparing Attributes Relationships between Mutual Datasets

Let us assign the weights: Equivalence = 5, Hierarchy = 3, Unknown = 1
The table can be converted into a matrix as follows:
Let Row vector: D1, Column vector: D2

\[
\begin{bmatrix}
3 & 1 & 1 & 1 \\
1 & 5 & 1 & 1 \\
1 & 1 & 1 & 1
\end{bmatrix}
\]

We assumed that the length of each row vector and column vector is proportional to the importance of the association. It can be evaluated that attribute relationship on the basis of scores. The length of each row vector and column vector can be calculated as follows:
Lengths of each attributes (column vector) on Dataset D1 are as follows:
\(DSD1[Attr_m] = \sqrt{11}, DSD1[Attr_n] = \sqrt{27}, DSD1[Attr_o] = \sqrt{3}, DSD1[Attr_t] = \sqrt{3}\)

When attributes of D1 are arranged in order of length:
\(Attr_n > Attr_m > Attr_o = Attr_t\)

Lengths of each attributes (row vector) on Dataset D2 are as follows:
\(DSD2[Attr_m'] = \sqrt{12}, DSD2[Attr_n] = \sqrt{28}, DSD2[Attr_j] = \sqrt{4}\)

When attributes of D2 are arranged in order of length:
\(Attr_n > Attr_m' > Attr_j\)

[5] CONCLUSION
In this paper, the authors selected the attributes; based on the different levels of security i.e. Restricted, Confidential and Public from the User- defined Big Data and the Reference Big data. Also the concept of Artificial Neural Network to train the machine in order to predict the correct and accurate output, i.e. the attribute that needs to be secured. In future, one can plan to introduce an algorithm to secure the attributes which is classified on the different levels of security.

REFERENCES


Author[s] brief Introduction
[1] Satya Prakash, Shantanu Kumar, Sneha Shankar, Sridhar K R
   8th Semester, Department of ISE,
   SJB Institute of Technology,
   Kengeri, Bengaluru – 560060

[2] Dr. Mohan H S
   Professor and Head, Department of ISE,
   SJB Institute of Technology,
   Kengeri, Bengaluru – 560060

Corresponding Address –
Dr. Mohan H S
Professor and Head, Dept. of ISE,
SJB Institute of Technology
BGS Health and Education City,
Kengeri, Bengaluru – 560060
Mobile: +919986079014