Integrating Fuzzy Logic and Data Mining: Impact on Cyber Security

TAPASYA PATKI
Student
B.Tech. (CSE)
MSIT, New Delhi
tapasya_patki@yahoo.co.in

A.B. PATKI
Sr. Director
Dept. of Info Tech.
GOI, New Delhi
apatki@mit.gov.in

V. KUMAR
Head of Deptt.
Comp. Sc. & Engg.
MSIT, New Delhi
nangia1@hclinfinet.com

A.Q. ANSARI
College of Comp. Sc.
King Khalid Univ.
Saudi Arabia
aqansari@ieee.org

Abstract: Data mining is the search for significant patterns and trends in large databases. Fuzzy Logic, on the other hand, provides techniques for handling cognitive issues in the real world. The paper discusses the application of fuzzy logic techniques and data mining practices on Cyber Security. With the introduction of e-commerce and e-governance applications as well as activity boom in cyber cafes, the pressure is on cyber security monitoring. Although the stream is primarily associated with Computer/IT professionals, it is being widely explored by the business community.

Key-Words: Data Mining, Information Mining, Fuzzy Logic, Rough Sets, Cyber Security, CRISP-DM

1 Introduction
Data mining techniques have been in the industry for almost two decades. Data Mining is the process of automating knowledge discovery through useful trends & patterns. In query tools for DBMS, the end user makes an assumption about some relation amongst various factors i.e. different field attributes of records in database. In contrast, for Data Mining environment, a user is asking a data-mining system to discover the most influential factors. A data-mining tool tries to discover relationship & hidden patterns that may not always be obvious. Data mining is neither data warehousing with SQL query reporting nor Online Analytical Processing (OLAP) / data visualization. Most commonly used data mining algorithms can be classified into two groups based on the philosophy of modeling i.e. algorithms using classical techniques & those deploying next generation methodologies [1-2]. While the classical techniques include statistics, neighborhood and clustering methods, the next generation techniques focus on principles using decision trees, artificial neural networks and rule-based systems. Model building process is central to data mining and representative model based on an existing data set has proven useful for understanding trends, patterns & correlation, as well as for predictions based on historical outcomes.

2 Industrial Scenario
With the increasing computing power and improved data collection & Management facilities, Chief Information Officers (CIOs) are concentrating on building data warehouses. Many DBMS query vendors are now offering data mining components with their software. A vendor driven trend of integrating data mining into the database is seen e.g. Oracle 9i (Darwin team works for the DB group, not applications), IBM Intelligent Miner V&R1, NCR Teraminer. While this Database-Mining Integration trend is a stopgap arrangement to provide one stop shopping, it is limited to analytics provided by vendor and hence other applications might not be able to access mining functionality. Thus, bundling of data mining software with DBMS will be identical to Internet Explorer tied up with Microsoft’s OS. Many established organizations like NCR, Diamler Chrysler are supporting Cross Industry Standard Process for Data Mining. The CRISP-DM project has developed a tool-neutral Data Mining process model to make large data mining projects faster, cheaper, more reliable and more manageable [3]. The life cycle of a data mining project consists of six phases viz. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The sequence of phases is not rigid and shifting to and fro between different phases is also needed, depending upon the outcome of the particular phase. A study of the CRISP-DM 1.0 step-by-step data mining guide released by the CRISP-DM Consortium in August 2000 reveals that business understanding and data understanding phases are usually human centered and only little routine automation can be achieved there.

3 Fuzzy Logic in Data Mining
CRISP-DM consortium type of efforts is leading to growing industrial interest in the data mining. It has
emphasized on cognitive aspects of business process at large. With the globalization of business practices, the efforts to supplement the consortium approach with adequate level of technologically supported products (hardware / software) and processes is the need of the hour to successfully deploy various phases mentioned in the standards. These phases are useful in defining

i) Goals of the knowledge discovery project
ii) Estimate potential benefits
iii) Identify and collect necessary data (including background domain and meta Knowledge)

In these phases, fuzzy set methods can be used to formulate, the background domain knowledge in vague terms. A fuzzy solution is not only judged for its accuracy, but also for its simplicity & readability. The salient highlights of fuzzy system of particular relevance to data mining are [4]

i) There are only few fuzzy rules in the rule base
ii) There are only few variables used in each rule
iii) No linguistic label is represented by more than one fuzzy set
iv) Fuzzy Logic rule induction can handle noise & uncertainty in the data values

4 Fuzzy Logic Based Algorithm, Implementation Issues & Software Development Considerations

Fig. 1 depicts a fuzzy logic based algorithm for data mining [4]. We illustrate various steps of the algorithm with an example. Consider the product sale for “Digital Calculator”. The following SQL query gives raw data from the database maintained at a stationary shop located in the school area in a residential locality. The shop is one amongst the chain of such stores owned by a single agency in various cities. Using a SQL query we have the primary data.

```
SELECT name, sex, age
FROM customer
WHERE place= "Lajpat Nagar" and item = "Digital Calculator"
Assume that the retrieved data is as given below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arora</td>
<td>F</td>
<td>14</td>
</tr>
<tr>
<td>Gupta</td>
<td>F</td>
<td>17</td>
</tr>
<tr>
<td>Joshi</td>
<td>F</td>
<td>12</td>
</tr>
<tr>
<td>Marwah</td>
<td>F</td>
<td>15</td>
</tr>
<tr>
<td>Radhe Shyam</td>
<td>M</td>
<td>30</td>
</tr>
</tbody>
</table>
```

**Step 1:** Define Fuzzy quantifications for all the variables. These could be triangular, trapezoidal or S / Pi Curve depending upon the nature of data and the choice of data mining analyst.

We choose primary Fuzzy quantifications as young, middle_age and old. Also we use hedges like “very” for young and old. We define the very_young and very_old membership functions instead of using Concentration (square) operations in typical fuzzy set texts. Thus, we have five fuzzy quantifications.

i) very_young, ii) young iii) middle_age iv) old v) very_old

**Step 2:** Replace every numerical value by a Fuzzy quantification name. Value is replaced by the fuzzy quantification whose membership grade is the highest. We build a symbol table using the membership function

<table>
<thead>
<tr>
<th>Name</th>
<th>Sex</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arora</td>
<td>F</td>
<td>young</td>
</tr>
<tr>
<td>Gupta</td>
<td>F</td>
<td>young</td>
</tr>
<tr>
<td>Joshi</td>
<td>F</td>
<td>young</td>
</tr>
<tr>
<td>Marwah</td>
<td>F</td>
<td>young</td>
</tr>
<tr>
<td>Radhe Shyam</td>
<td>M</td>
<td>middle_age</td>
</tr>
</tbody>
</table>

Let the class be denoted as CALCULATOR_SALES, which is used as output variable and for this output variable, let the fuzzy quantifications be “high”, “medium” and “low”.

**Step 3:** Consider every row in the fuzzy symbol table as a fuzzy rule. Apply minimization technique to reduce the number of fuzzy rules. Fuzzy rule Minimization is similar to minimization techniques used for a switching function in Boolean algebra [4].

Fuzzy Sum-Of-Products Expressions

CALCULATOR_SALES.high =

\[(\text{sex.F})(\text{age.young}) + \text{(sex.F)}(\text{age.young}) + \text{(sex.F)}(\text{age.young}) + \text{(sex.F)}(\text{age.young}) + \text{(sex.M)}(\text{age.middle_age})\]

**Step 4:** Simplify the Fuzzy Sum-Of-Products Expressions by eliminating the redundant products and formulate Fuzzy rules from Fuzzy Expressions. Simplifying the above fuzzy expression, we get

CALCULATOR_SALES.high =

\[(\text{sex.F})(\text{age.young}) + \text{(sex.M)}(\text{age.middle_age})\]

The fuzzy expression, \((\text{sex.F})(\text{age.young})\), should be weighted more than the fuzzy expression \((\text{sex.M})(\text{age.middle_age})\), because the former accounts for four records out of five records (i.e.
80%), while the latter for only 20 % (i.e. only one record out of five). Hence, CALCULATOR_SALES.high = (sex.F)(age.young) Thus, we have a fuzzy rule
IF SEX IS F AND AGE IS YOUNG THEN CALCULATOR_SALES IS HIGH.
resulting into a N-input-One-output Fuzzy System. In case of a relational database, an SQL query can be entered to retrieve data/records. Providing add on module for SQL to implement a fuzzy data mining algorithm is a makeshift arrangement. Presently, fuzzy logic based front-end interface implementations for commercial database software systems have not been deployed. We have implemented a Fuzzy logic Operating System support software using C++ for FUZOS© and exploratory extensions for fuzzy data mining are under prototype module development and testing. In order to facilitate web mining using fuzzy logic based Java applets, potentials of Java platform vis-à-vis Visual Prolog are being explored.

5 From Data Mining to Information Mining - Application in Cyber Security
The general hesitation for non acceptance of “data-mining” philosophy in the industry is a matter of serious concern and has adverse impact on “web mining” in the context of Information Technology (IT) for masses. The standard definition of knowledge discovery and data mining only speaks of discovery in data. Usually a minimal requirement is a relational database. Most methods (including decision trees and neural networks) even demand input as a single uniform table i.e. a set of tuples of attribute values. The learning procedure in case of neural networks is not easily comprehensible and the user does not get insight into the domain from where the data comes. Further, it cannot handle heterogeneous data i.e. mix of image, sound, textual data or even textual descriptions in scanned files. Although such transformations of data into structured tables using feature extraction are recommended practices, these are inadequate. The situation is identical to the scenario when COBOL programs using file systems for pseudo-data base applications were used prior to availability of relational data base schemes introduced by E.F.Codd in 1970s.
Data Mining standards like CRISP-DM have emphasized upon the cognitive aspects of business process shifting the focus from data mining towards information mining, which encompasses the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in heterogeneous information sources. Today, for want of commercial software packages, consultants with their in house proprietary software tools are performing many of the heterogeneous data handling tasks.

6 Avenues for Information Mining in Cyber Security
With the increasing usage of e-commerce applications along with wide spread deployment of computer networks in today’s society, cyber security has assumed high priority in Government, industrial and business forums. Cyber legislation in the form of Information Technology (IT) Act 2000 is operational in India. Technology development trends to support Community Informatics and Cyber Civilization vis-à-vis Cyber Crimes information systems have been analyzed for deploying fuzzy logic solutions [5,6]. Community Information Centers (CICs) and large-scale expansion of cyber café infrastructure (with wideband connectivity), has opened up several new information-mining activities in “real time” domain where the present day data mining techniques are proving ill equipped. In order to give an idea of information mining application, we describe a typical fuzzy intrusion detection scenario to investigate vulnerabilities of computer network. Intrusion detection focuses on
i) Misuse detection and
ii) Anomaly detection
Here the heterogeneous data is generated both from network engineering monitoring measurements (hardware devices security related feature) as well as software audit data produced by network security manager. In such applications, non-stationary processes generate the data streams and hence the problem falls in the category of real time domain. It is felt that an incremental approach to mining non-stationary data streams when deployed here is likely to increase the average classification rate of real time data mining systems by reducing the number of times a completely new model is generated. More or less, we need to update the existing model data instead of constructing a completely new model, as long as no “concept drift” is detected. Lower bound and upper bound methods of ‘rough set’ theory are useful, in determining the concept drift [7].
Incremental approach is suitable for ‘real time’ data mining applications as it reduces processing time
by performing minimal changes in the current structure of the classification model. We have analyzed the applications of soft computing methodologies and currently extending scope of chaos theory including fractal mathematics as fundamental building blocks for information mining. A typical fuzzy logic rule could be

IF the number of different destination IP addresses during the last few seconds was high

THEN an unusual situation exists.

Here, the terms few and high are fuzzy terms in antecedent portion of the rule and unusual is the fuzzy term in consequent portion of the rule. In order to illustrate the problem, we give a fuzzy set for Destination ports as indicated in Fig. 2 using a triangular membership function representation. Internet Service Providers (ISPs) are the agencies, which provide relevant support to the cyber cafes, Internet Kiosks and similar small and medium enterprises. A large number of users are constantly and regularly using Internet for their variety of applications from geographically dispersed locations. The profile of cyber cafe usage, types of browsers used, time of day vis-à-vis physical location are amongst the primary criteria for assessing the potentials of cyber crimes through the ‘floating user’ population of susceptible cyber cafes. This data is useful in deciding a priori probability calculation of the concerned potential cyber crime threats. Although huge data is readily available, it is not being put to use for want of adequate software support for analysis as well as due to non-availability of handheld portable gadgets for monitoring. Even Network security vendors who are supporting sophisticated network security operations and who themselves grew out of service providers role, like Juniper Networks do not have adequate data mining support in their latest operating system like JUNOS version 7.1 in spite of the fact that several gigabytes of data is analyzed through simple statistical measures. Data Mining, which was in the past considered as an off line activity, is assuming new dimension as an inherently on line task for Internet security monitoring. Here, the handheld portable gadgets with the law enforcing agencies / cyber forensic analysts will have provision to download small information from ISP web server for effective monitoring in real time mode. Using these handheld gadgets, with the downloaded data, which is based on Shanon’s information theory, is useful for cyber crime prevention. According to Shanon’s information theory, if p is the probability of occurrence of a message, then the information gained from the message, I, is given by

\[ I = \log_2 \left( \frac{1}{p} \right) = - \log_2 p \]  

(1)

Suppose we are trying to categorize cyber cafe locations that fall into ‘risky’ and ‘safe’ classes. This is not a permanent classification. It is changing continuously depending on the time of day, location of cafe (in isolated area, in crowded area like railway stations), Internet browser available at the cyber cafe (Internet Explorer, Netscape, etc.) and hence calls for ‘real-time’ categorization. The data available with ISPs is useful for deciding a priori probabilities and helps in assisting the monitoring personnel by on the fly downloading on to their handheld devices. Let us assume we had 1000 cafes in a city 800 of which were safe and 200 were risky. We would have a training set with 1000 records containing some set of attributes of these cafes (location, browser type, time of day, sex of users, age of users, family status like affluent / middle class in the cyber cafe surroundings etc.) as well as information whether a cafe was in risky zone or in a safe category. Safe cafes are considered as positive examples (p) and risky cafes are considered as negative examples (n).

If we had downloaded information from ISP’s web server of a cafe and had to try to predict whether that cafe would be in a safe or risky category during the monitoring activity (in real time domain), how much information would be contained in a correct answer. Using Shannon’s information theory, we have

\[ I \left( \frac{p}{p+n}, \frac{n}{p+n} \right) = - \left( \frac{p}{p+n} \right) \log_2 \left( \frac{p}{p+n} \right) - \left( \frac{n}{p+n} \right) \log_2 \left( \frac{n}{p+n} \right) \]  

(2)

The chance that any single factor (e.g. browser type, age of user, location) would completely divide the group into those who are risky or safe is small. We need to measure how much information we still need after the test. Let any attribute A, which has V distinct values that divides the data set into V subsets. Each resulting subset of the training data has its own characterization of p and n outcomes. On average, after testing attribute A, we still need

Remainder (A) = \sum (p_i + n_i) I \left( \frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i} \right) \]  

(3)

bits of information, where, \( i \) varies from 1 to \( V \), the number of discrete values that attribute A can take. The Information gain is defined as the difference between the information needed before the attribute test and the remainder.

\[ \text{Gain (A)} = I \left( \frac{p}{p+n}, \frac{n}{p+n} \right) - \text{Remainder (A)} \]  

(4)

Consider the attributes and Values as below.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>V=2</td>
<td>Residential, Crowded</td>
</tr>
<tr>
<td>Usage Time</td>
<td>V=3</td>
<td>Morning, Day, Night</td>
</tr>
</tbody>
</table>

Suppose that cafes in residential areas are found safe 90% of the time and those in crowded areas 70% and that our café set is made up half of residential and half of crowded locality. Our question is how much information gain would we get simply by testing whether a café is located in residential or crowded area. Using equation (4)

\[
\text{Gain (Location)} = 1 - [0.5 I (450/500, 50/500) + 0.5 I (350/500, 150/500)] = 0.325
\]

Suppose, that we had grouped the café’s usage habits into 3 groups viz. late night hours, daytime and morning college hours. Further assume that these were evenly split between all the cafes. When we look at the cyber crime rates, we see those first groups are safe at 50 %, the second at 90% and the third at 100%. The information we would gain by testing this attribute is:

\[
\text{Gain (Usage)} = 1 - [(0.33) I (166/333, 166/333) + (0.33) I (300/333, 33/333) + (0.33) I (1,1)] = 0.179
\]

As is clear, the first case i.e. location gives us much more information gain (0.325 bits) than the second (i.e. 0.179). So when the monitoring surveillance personnel download such information from the ISPs web server on their handheld gadget, it helps them to physically visit and prevent the cyber crime attitude.

7 Conclusions

In knowledge discovery and data mining, the present methodology is to focus on purely data-driven approaches. However, to arrive at really useful results, we must take non-alphanumeric information also into consideration. This article opens up a discussion on the issues of transition from data mining (homogeneous) to information mining (heterogeneous). It also describes applications of fuzzy logic for data mining and extends its scope for information mining in an integrated manner with techniques from rough set methodologies and chaos theory. It further reflects on the combined usage of Data Mining and Fuzzy Logic based techniques for dealing with Cyber Security issues in the present era by introducing of cyber crime prevention practices.

Acknowledgements

Discussions with industry professionals have helped in understanding hesitation of industrial houses to embrace data mining philosophy. Author from DIT wishes to thank several useful interactions he had with Shri S. Lakshminarayanan, Secretary, Inter-State Council Secretariat Government of India.

References:

Simplify Fuzzy Sum-Of-Products Expressions

Apply Minimization Techniques on fuzzy Expressions

Replace Numerical variables by Fuzzy quantification using Membership functions

Define fuzzy quantifications for Problem set

Obtain Fuzzy Rules

Fig 1. Flow Chart for a Fuzzy Logic Based Algorithm for Data Mining Rules

Low

Medium High

Destination Ports

0 5 10 15 20 25 30

Fig 2. Fuzzy Membership Function: Variable HIGH in Section 4