Integrating Fuzzy Logic and Data Mining: Impact on Cyber Security

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

Tapasya Patki
Student
B.Tech. (CSE), MSIT
New Delhi, India
tapasya_patki@yahoo.co.in

A. B. Patki
Senior Director
Dept. of Info. Tech.
New Delhi, India
apatki@mit.gov.in

V. Kumar
Head of Deptt.(CSE)
MSIT, GGSIP Univ.
New Delhi, India
hodcse@msit.in

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 in 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 and corporate legal community. Existing data mining solutions are not directly adaptable to support E-Discovery legal compliance process. We discuss and illustrate the scope of fuzzy logic to circumvent some problems in the cyber crime domain.

1. Introduction

Data Mining is the process of automating knowledge discovery through useful trends and 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. 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 and 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 useful for understanding trends, patterns and correlation, as well as for predictions based on historical outcomes. Federal Bureau of Investigations (FBI) Cyber division created in 2002 coordinates international cyber crime investigations. In the coming years, high profile IT security activity in terms of risk management and response is likely to integrate IT and legal functions. Electronic discovery requests are gaining significance and searching for responsive data has caused concerns leading to conflicts of interest for IT department and legal personnel. For the multi-faceted projects, role of electronic evidence experts, assume growing demand to decide in favor of outsourcing or otherwise [8]. In the next section we describe the prevailing industrial scenario to get a feel of E-Discovery activity and potential problems. In the subsequent sections 3 and 4, we illustrate the need for moving towards integration of fuzzy logic and data mining to support IT-legal framework to face the challenges of E-Discovery process. We have illustrated the applicability of fuzzy logic through examples akin to functions of IT security division.

2. Industrial Scenario

Traditional discovery rules to contend the paper documents have undergone transformations in the digital evidence discovery process. Federal Rules of Civil Procedure (FRCP) amendments expand the scope for electronic discovery services ranging from data gathering, media restoration through data processing, for legal evidence. This additional compliance burden on IT departments necessitates the extension of concepts of data mining practices to improve its suitability during litigation, investigations and regulatory compliance. We see the limitations of data warehousing approaches since these were primarily focused for IT as end user and not IT as compliant end user. With the increasing computing power and improved data collection and management facilities, Chief Information Officers (CIOs) are concentrating on building data warehouses. A vendor driven trend of integrating data mining into the data base is seen e.g. Oracle 9i (Darwin team works for the DB group, not applications), IBM Intelligent Miner V&R1, and NCR Teraminer. While the Database-Mining Integration trend is an arrangement to provide one stop shopping, it is limited to analytics provided by vendor and

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

hence other applications (e.g. for E-Discovery compliance for amendments to FRCP) might not be able to access mining functionality effectively. Thus, bundling of data mining software with DBMS will be identical to Internet Explorer tied up with Microsoft's Windows 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. We need to integrate these phases dynamically to incorporate the scope of FRCP amendments at each phase to report the feedback so that the turn around time for compliance reporting is minimal. A fuzzy rule based system to support cognitive aspects is recommended since the existing software solutions do not incorporate cognitive aspects to support FRCP amendments. 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. It is highly recommended that scope of CRISP-DM be augmented to encompass the FRCP amendments to provide vendor neutral seamless solutions.

3. Fuzzy Logic in Data Mining and Information Mining

CRISP-DM consortium effort is emphasizing cognitive aspects of business process. Cognitive aspects are also associated with FRCP amendments. 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)

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. These limitations of the existing practices are the bottlenecks in deploying off-the-shelf data mining products for addressing FRCP amendments. Further, it cannot handle heterogeneous data i.e. mix of image, sound, textual data or even textual descriptions in scanned files. Although data transformation into structured tables using feature extraction is recommended, it is inadequate to address cognitive issues. 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. Such activities are not upward scalable for supporting E-Discovery related work in IT departments for FRCP amendment compliance. We are in the transition from data mining to information mining.

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 and 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 and uncertainty in the data values

4. Fuzzy Logic Based Algorithm, Implementation and 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.

Name	Sex	Age
Arora	F	14
Gupta	F	17
Joshi	F	12
Marwah	F	15
Radhe Shyam	M	30

Step 1: Define Fuzzy Membership Functions for all the variables. These could be triangular, trapezoidal or S / Pi. 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

Name	Sex	Age
Arora	F	young
Gupta	F	young
Joshi	F	young
Marwah	F	young
Radhe Shyam	M	middle_age

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 Fuzzy Rule Minimization similar to minimization techniques used for a switching function in Boolean algebra [4] to reduce the rule base size.

Fuzzy Sum-Of-Products Expressions

CALCULATOR_SALES.high = (sex.F)(age.young) + (sex.F)(age.young) + (sex.F)(age.young) + (sex.F)(age.young) + (sex.M)(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 = (sex.F)(age.young) + (sex.M)(age.middle_age)

The fuzzy expression, (sex.F)(age.young), should be weighted more than the fuzzy expression (sex.M)(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). So 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. 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. Information Mining in Cyber Security and E-Discovery

With the increasing usage of e-commerce applications along with widespread deployment of computer networks in today's society, cyber security has assumed high priority in Government, industrial and business forums. 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]. In order to give an idea of information mining application in real time domain, 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. 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].

A large number of users are constantly and regularly using Internet for variety of applications from geographically dispersed locations. The profile of cyber café 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. Incremental approach using fuzzy rule set is suitable for 'real time' information mining to reduce processing time by performing minimal changes in the current structure of the classification model. 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.

Although huge data is readily available with Internet Service Providers (ISPs), it is not being effectively used due to inadequate software support for analysis as well as due to non-availability of handheld portable gadgets for monitoring. 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. 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. Using these hand held gadgets with the downloaded data, assessment using Shanon's information theory is undertaken for cyber crime prevention. In the context of E-Discovery, it is necessary to focus on parameters that provide maximum evidence i.e. we have to exclude the factors where information gain is minimal. We illustrate the concept through an example.

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

$$\mathbf{I = \log_2 (1/p) = - \log_2 p} \qquad \mathbf{(1)}$$

Suppose we are trying to categorize cyber café locations that fall into 'risky' and 'safe' classes. It is changing continuously depending on the time of day, location of café (in isolated area, in crowded areas like railway stations), Internet browser available at the cyber café (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 police patrolling 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 café surroundings etc.) as well as information whether a café 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). Downloaded information from ISP's web server of a café is used to predict whether that café would be in a safe or risky category during the monitoring activity (in real time domain). Using Shannon's information theory, we have

$$\mathbf{I (p/p+n,n/p+n) = - (p/(p+n)) \log(p/(p+n)) - (n/(p+n)) \log(n/(p+n))} \qquad \mathbf{(2)}$$

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 it's own characterization of p and n outcomes. On average, after testing attribute A, we still need

$$\mathbf{Remainder\ (A) = \Sigma\ (p_i + n_i\ I\ (p_i/p_i+n_i,\ n_i/p_i+n_i)} \qquad \mathbf{(3)}$$

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.

$$\mathbf{Gain\ (A) = I\ (p/p+n,\ n/p+n) - Remainder\ (A)} \qquad \mathbf{(4)}$$

Consider the attributes and values as below.

Attribute	Values	Description
Location	V=2	Residential, Crowded
Usage Time	V=3	Morning, Day, Night

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)

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

Suppose, that we had grouped the café's *usage habits* into 3 groups viz. late night hours, daytime and morning college hours and 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:

$$\mathbf{Gain\ (Usage) = 1 - [(0.33)\ I\ (166/333,166/333) + (0.33)\ I\ (300/333,\ 33/333) + (0.33)\ I\ (1,0)] = 0.179}$$

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 ISP's web server on their handheld gadget, it helps them to physically visit and prevent the cyber crime attitude. The methodology

suggested is useful for supporting E-Discovery activity in the organization especially to know on which topics the information gain is minimal.

7. Conclusions

In knowledge discovery and data mining, the present methodology is to focus on purely data-driven approaches. This article opens up a discussion on the issues of transition from data mining (homogeneous) to information mining (heterogeneous) using fuzzy logic. 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 cyber crime prevention practices. The scope of E-Discovery and integration of such concepts have been brought out.

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. Technical inputs given by Mr. Mahesh Kulkarni, Group Coordinator, CDAC, Pune, India have been helpful in improving the coverage of the paper.

References:

[1] Pieter Adrians, Dolf Zantinge, - Data Mining, Syllogic Press, 2003 Pearson Education

[2] Abbass H.A., Sarker R.A., Newton C.S. – Data Mining: A heuristic Approach, Idea Group Publishing, 2002

[3] Cross Industry Standard process for Data Mining – http://www.crisp-dm.org/

[4] Patki A.B. Impact of fuzzy Logic on Data Mining Practices, Seminar on Data Warehousing, Data Mining and Business Intelligence, March 12, 2005, MSIT, New Delhi

[5] Patki A.B., Kulkarni M.D., Subramanian S., Patki Dhanvanti D. - Technology Development Trends for Cyber Civilization, Proceedings of 2003 International Conference on Cyberworlds, (CW 2003), December 3-5, 2003, Singapore, IEEE Computer Society pp.40-45

[6] Patki A.B., Kulkarni M.D., Patki Dhanvanti D. - Software Development Paradigms for Community Informatics- Technological Aspects, Proceedings of 5th International Conference on Information Technology in Regional Areas (ITiRA), December 15-17, 2003, Caloundra, Queensland, Australia

[7] Patki A.B., Raghunathan G.V., Ghosh S., Sivasubramanian S. - Towards Rough Set Based Concept Modeler, WSC4, Fourth On-line World Conference on Soft Computing in Industrial Applications, Sept. 24-30, 1999, Nagoya, Japan.

[8] Linda Kish, Knowing Your Limits: When to Outsource E- Discovery, E-Discovery Advisor Magazine, Web Edition 2006, Week 11, doc # 17590

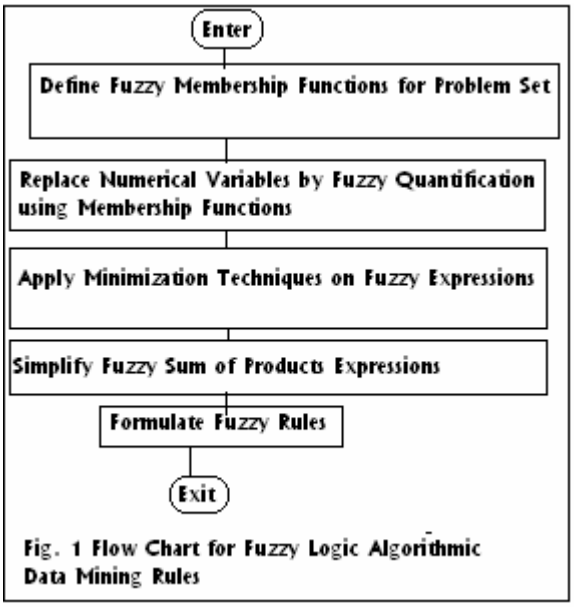

Fig. 1 Flow Chart for Fuzzy Logic Algorithmic Data Mining Rules

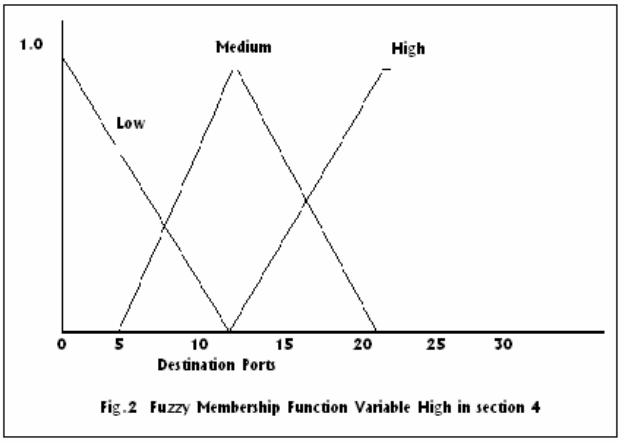

Fig.2 Fuzzy Membership Function Variable High in section 4