UNIVERSITY OF CANBERRA

This is the published version of this work:

Mohammadian, M. (2008). Classification of Data Based on a Fuzzy Logic System. In M. Mohammadian (Ed.), *International Conference on Computational Intelligence for modelling , Control and Automation,* pp. 1288-1293. United States: IEEE, Institute of Electrical and Electronics Engineers.

This file was downloaded from:
https://researchprofiles.canberra.edu.au/en/publications/classification-of-data-based-on-a-fuzzy-logic-system

©2008 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

Notice:
The published version is reproduced here in accordance with the publisher's archiving policy 2008.

Classification of Data Based on a Fuzzy Logic System

Masoud Mohammadian
Faculty of Information Science and Engineering
University of Canberra
Canberra, ACT, Australia
masoud.mohammadian@canberra.edu.au

Abstract

Data security and privacy are very important issues in the success of a business operation. Implementing and applying policies related to data security and privacy therefore has become one of the core and important activities for large organizations. Data classification process allows companies to organize their data according to their needs. This process can be laborious in large organizations with significant content to evaluate and categorize. Using a data classification process organizations can identify and apply appropriate setting and polices such as private access control and encryption requirements only to the relevant data thereby saving time and processing power. This paper explores the use of fuzzy logic in classification of data and suggests a method that can determine requirements for data security and privacy in an organization based on organizations needs and government policies imposed on data. A Case study is considered to present the effectiveness of the proposed method.

1. Introduction

Organizations understand the value of their data and need for adequate data protection services. Data security, availability, privacy and integrity are very important issues in the success of an organization operation. Data security and privacy policies in organizations are governed by business requirements and government regulations. These requirements demand not only for data security but also for data accessibility and integrity.

Implementing data security using data encryption solutions remain at the forefront for data security. Data encryption algorithms are implemented to protect the actual data. However the issue of keys and overall process of data encryption process remains complex. The best data encryption solutions are those that balance information protection with on-demand access to that encrypted data. Using a data classification process organizations can identify and encrypt only the relevant data thereby saving time and processing power. Without data classification organizations using encryption process would simply encrypt everything and consequently impacting users more than necessary [1, 2].

Data classification is essential and can assist organizations with their data security, privacy and accessibility needs. Such a classification process needs to be able to determine the value, sensitivity, privacy, government regulations and corporate strategic objectives. However data classification is a major difficulty for many organizations as it is an expensive and time consuming task. This paper explores the use of fuzzy logic for classification of data and suggests a method that can determine requirements for classification of organization's data for security and privacy based on organizations needs and government policies imposed on data.

2. Data classification in large organizations

An understanding of the nature of the organization and its activities and usage of data is required to assist in data classification. Based on data usage processes can be identified to protect the data. Data can be classified to be public, private, private but not mission critical, mission critical, vital, secret, or top secret [1, 2]. There is no exact and firm rule on the level of classifications however issues such as data type, level of data sensitivity and corporate objective and regulatory rules are a few to consider. Such data classification will be different for different organizations based on policies of each

978-0-7695-3514-2/08 $25.00 © 2008 IEEE
DOI 10.1109/CIMCA.2008.232

IEEE computer society

organization and government regulatory polices. Policies are expressed in human understandable language and are vague and difficult to represent formally. The excessive gap between precision of classic logic and imprecision and vagueness in definition of polices creates difficulty in representing this policies in formal logic. Fuzzy Logic [4, 7, 8] has been found to be useful in its ability to handle vagueness. In this paper a data classification method based on fuzzy logic is presented to determine data classification levels for data in an organization. The level of sensitivity and corporate objective and regulatory rules are determined using this classification method.

3. Fuzzy logic for data classification in large organizations

Classification levels could divide data into classes such as "top secret", "secret", "confidential", "mission critical", "not critical", "private but not top secret", and "public". Based on these class categories the business processes and individuals that access and use the data and the level of encryption can be identified. The users can be categorized to determine access to any of these data classes. Such a user classification can classify user into classes such "very high", "high", "medium" or "low" level access authorization users. The need for encryption level of the data can also be determined to be high, medium, zero (i.e not necessary).

To classify data with minimal resources impact and without needing to re-design databases one option is to add extra information to each data item by adding meta-data information to the attributes of each entity in relational-data bases and domains (concepts) in classes in object-oriented databases.

These meta-data information could be the value or degree of security, privacy or other related policies for that data item. Below a simple relational database is used for demonstration example. Consider the following entities in a relational database system:

Customer (CustomerID, Name, Address, TelNo, E-mail)
Product (ProductID, Name, Size, Color, Price)
Supplier (SupplierID, Name, Address, TelNo, FaxNo, E-mail)
Order (OrderID, CustometID, ProductID, SupplierID, OrderDate, Quantity)

Meta-data values can then be used for adaptation and implementation of data classification of data in databases for an organization. The meta-data values can be obtained from the knowledge workers of the organization based on organization policies, procedure and business rules as well as government requirements for data privacy and security. For example table 1 show the metadata values related to security attributes of table Customer based on organization's security policy and government security and privacy policy. The values are in the range of 0 to 70, where zero indicates a meta-data for a data item that is public and 70 indicates the meta-data for a data item that is top secret. Note that other values are also possible. Now assume that the following domain meta-data values for these linguistic variable, **TP** = top secret, **SE** = "secret", **CO** ="confidential", **MC** = "mission critical", **NC** = "not critical", **PR** = "private but not top secret", **PU** = "Public". The values related to linguistic vraiables are: **TP** = [58,..,70], **SE** = [48,..,60], **CO** =[37,..,50], **MC** = [28,..,40], **NC** = [16,..,30], **PR** = [8,..,20], **PU** = [0,..,10]. Based on the metadata value for each attribute the membership of that attribute to each linguistic variable can be calculated. Fuzzy sets can be used to represent the data security classifications (e.g. Data security classification levels: **TP** = top secret, **SE** = "secret", **CO** ="confidential", **MC** = "mission critical", **NC** = "not critical", **PR** = "private but not top secret", **PU** = "Public"). The meta-data values for the attributes of table Customers is given in Table 1:

Customer Table	Meta-data Value base on organization policy	Meta-data Value base on government regulatory policy
CustomerID	68	39
Name	64	70
Address	30	60
TelNo	44	68
E-mail	67	69

Table 1. Metadata values for table Customer

The degree of membership value of the attribute CustomerID to fuzzy set data classification based on meta-data from Table 1 is given in Table 2. Now that the data can be classified and categorized into fuzzy sets (with membership value), a process for determining precise actions to be applied must be developed. This task

involves writing a rule set that provides an action for any data classification that could possibly exist. The formation of the rule set is comparable to that of an expert system, except that the rules incorporate linguistic variables with which human are comfortable. The use of fuzzy sets allows rules to be derived easily based mostly on the organizations and government regulatory polices.

TP	SE	CO	MC	NC	PR	PU
0.8	0	0	0	0	0	0

(a)

TP	SE	CO	MC	NC	PR	PU
0	0	**0.3**	**0.16**	0	0	0

(b)

Table 2. Fuzzy membership of metadata value of CustomerID based on: (a) Organization policy, (b) government regulatory policy

Fuzzy If-Then rules then can be built to act on data for further actions on the data based on corporate strategy for data security, data privacy as well as government regulations for data security and data privacy. The fuzzy rules could be of the form [8, 9]:

If [a_1 is ($\mathbf{A}_1$) and a_2 is ($\mathbf{A}_2$) and ...] Then [b_1 is ($\mathbf{B}_1$) ALSO b_2 is ($\mathbf{B}_2$) ALSO....]

Where $\mathbf{A}_i$ is the fuzzy set characterizing the respective decision variables (in this case the data that is classified based on meta-data value) and $\mathbf{B}_i$ is the fuzzy set characterizing the action variables. Although all possible conditions in the physical system seems imposing at first, the incorporation of fuzzy terms into the rules makes the development much easier. The fuzzy rules (**A**, **B**) associate an output fuzzy set **B** of the action values with an input fuzzy set **A** of input-variable values. We write fuzzy rules as antecedent-consequent pairs of If-Then statements. For example:

IF *Organizational_Security_Classification* is **TopSecret** and *Government_Security_Classification* is **Confidential Then** *Level of Encryption required is* **High**

The overall fuzzy output is derived by applying a decision formula [9] such as the "max" operation to the qualified fuzzy outputs each of which is equal to the minimum of the firing strength and the output membership function for each rule. In this paper the following decision formula is used to deducted the required output:

$$Output = \frac{\sum_{i=1}^{n} \alpha_i \mu_i}{\alpha_i} \qquad (1)$$

Where a_i the upper bound value of the fuzzy set *i* and μ_i is the membership value of the fuzzy set *i*. The process of data retrieval based on fuzzy logic.

4. Case study and results

A organization wishes to classify their data x_k , *k = 1,..,p* which are stored in their database. There exist government and regulatory polices G_i , $i = 1,...,n$ and organizational policies P_j , $j = 1,...,m$. Assume now that every data item x_k is evaluated and has been given meta-data values representing values or degrees of government regulatory security, privacy G_i and organizational policies P_j for that data item. The G_i and P_j values are in the range of 0 to 70, where zero indicates the meta-data for a data item that is public and 70 indicates the meta-data for a data item that is top secret. Assume that the linguistic terms describing the meta-data for each data item x_k in the above database are set to be **TP** = top secret, **SE** = "secret", **CO** ="confidential", **MC** = "mission critical", **NC** = "not critical", **PR** = "private but not top secret", **PU** = "Public" with the following values for each linguistic variable **TP** = [58,..,70], **SE** = [48,..,60], **CO** =[37,..,50], **MC** = [28,..,40], **NC** = [16,..,30], **PR** = [8,..,20], **PU** = [0,..,10]. The meta-data value given to a data item x_k concerning government and regulatory polices G_i is denoted by a_k and that concerning the organizational related policies P_j is denoted by b_k . Using these meta-data values we construct discrete fuzzy sets G_i and P_j on the set of alternatives A_{alt} such that:

$$G_j = \{(x_a, a_{1i}),...,(x_p, a_{pi})\}, \quad i = 1,...n$$
$$P_j = \{(x_a, b_{1j}),...,(x_p, a_{pj})\}, \quad j = 1,...m$$

Then a decision formula can be constructed [7] given by the following formula:

$$D = G_1 \cap .... G_n \cap P \cap P_n \quad \text{where}$$

$$\mu_k = \min(a_{k1}..., a_{kn}, b_{k1},..., b_{km}),\ k = 1,..., p$$

The polices with the highest membership grade among $\mu_1,..., \mu_p$ will be considered the policy to be applied for the required level of encryption. Now assume that there are government and regulatory polices and organizational polices for a given data item $x_i, i = 1,..,2$ form the set of alternatives $A_{alt} = \{x_1, x_2\}$ where the membership of x_1 in fuzzy set of government polices is $\mu_{CO} = 0.3$ and $\mu_{MC} = 0.16.$ The membership of x_2 in fuzzy set of organization polices is $\mu_{TP} = 0.8.$

	TS	SE	CO	MC	NC	PR	PU
TS	HI	HI	HI	HI	HI	HI	HI
SE	HI	HI	HI	HI	HI	HI	HI
CO	HI	HI	HI	HI	HI	HI	HI
MC	HI	HI	HI	HI	HI	HI	HI
NC	ME	ME	ME	ME	ME	ME	ZE
Pr	HI	HI	ME	ME	ME	ME	ME
PU	HI	HI	ME	ME	ZE	ME	ZE

Table 3. Fuzzy knowledge base for case study above

Now a decision formula can created $D = \{(x_1, 0.3), (x_2, 0.8)\}$. Now the policy x_2 has the highest membership value 0.8 hence this membership value will be applied to the consequence of the rules (see Table 3) that are triggered by the condition values of x_1 and x_2. The rules that were invoked were:

IF Organizational_Classification is **Top Secret** and Government_Classification is **Confidential Then** Level of Encryption required is **High**

IF Organizational_Classification is **Top Secret** and Government_Classification is **Mission Critical Then** Level of Encryption required is **High**

Finally the result will be calculated using the decision formula (see equation 1) and the the decision value 0.8 (obtained from the decision formula) is applied to the consequence of the above rules which means that the data item needs to be encrypted before it is stored in the database.

5. Conclusion

Data classification in organizations is a fundamental requirement for adequate information privacy and security. The consequences for not fully implementing a data classification scheme in organizations can be severe and costly in financial sense and organization's reputation. However many organizations do not have a classification scheme for their data. Development of such data classification can be very expensive. The cost of data classification is first to develop classification schema and then acquiring information from the personnel that are able to classify data. In many case large amount of data in organization are unclassified. Many medium size organizations classify all their data as being confidential and encrypted all their data although some of these data may not be confidential and may not require encryption. The proposed method in this paper provides a suitable data classification based on fuzzy logic.

Allocation of meta-data values for classification is not a simple or intuitive process. It requires a careful evaluation of data against a broad range of organization and regulatory polices which can make this process complex. But once data classification is accomplished the implementation of security and privacy can be performed successfully.

It should be noted that the proposed data classification method requires knowledgeable employees to recognize and classify data accordingly. A substantial effort is required to classify existing data and to continue to classify new data and re-classify some existing data. Future work in this area will explore the notion of time related to classified data based on the proposed fuzzy data classification to allow for re-classification of data after certain time. An example of such a situation is the case of a company earning announcement. Such data is time related. That means it is confidential until the company releases it's earning [3]. Another direction to improve security can be achieved by identification of a data owner and addition of this information to the fuzzy logic classification method.

Acknowledgement – Author wish to acknowledge the assistant and support provided to him during is visit to the Department of Electrical and Computer Engineering at the University of Toronto for this project by Professor Dimitrios Hatzinakos.

6. Reference

[1] J. Cline, "Growing pressure for data classification", http://www.computerworld.com/action/article.do?articleId=9014071&command=viewArticleBasic, Last accessed on 22/11/2008.

[2] R. Butterfield, "Data classification: A prerequisite to ILM", http://www.snwonline.com/implement/data_class_05-30-05.asp
Last accessed on 22/11/2007.

[3] R. Collette and M. Gentile, Overcoming Obstacles to Data Classification", http://www.computereconomics.com/article.cfm?id=1117 Last accessed on 22/11/2007

[4] L. A. Zadeh.,"Fuzzy sets", *Information and control*, Vol. 8. pp 338-352, 1965.

[5] D. L. Clark, "*Enterprise Security – The manager's defense guide*", Addison-Wesley, USA, 2003.

[6] G. McGraw, "*Software Security*", Addison-Wwesley, USA, 2006.

[7] H. H. Hosmer, "Using Fuzzy Logic to Represent Security Policies in the Multipolicy Paradigm", *ACM SIGSAC Review*, 1993.

[8] G. Bojadziev and M. Bojadziev, "*Fuzzy Logic for Business, Finance and Management*", 2nd Ed, World Sceintific, Singapore, 2007.

[9] B. Kosko, *Neural networks and fuzzy systems, a dynamic system*., Prentice-Hall: Englewood Cliff, USA, 1992.