ISSN 1330-3651(Print), ISSN 1848-6339 (Online)
UDC/UDK 004.056:519.226

DECISION SUPPORT BASED ON THE RISK ASSESSMENT OF INFORMATION SYSTEMS AND BAYESIAN LEARNING

Hrvoje Očevčić, Krešimir Nenadić, Krešimir Šolić

Original scientific paper
Risk protection has long been one of the main tasks of companies in a wide scope of business. From extensive range of risks the cyber-risks highlight as one of the most important. Cyber-risks are generated from hackers, malicious software, disgruntled employees, competitors, and many other sources both internal and external. Internal and external attacks on corporate assets and rapidly growing technology forced corporate management to conduct more appropriate awareness of the information security risks to information assets. The information security risk assessment, when performed correctly, can give corporate managers the information they need in order to understand and control the risks to their assets. The risks are in much more detail analysed in economic sectors, but in recent years there is increasing of risk assessment practice in the world of information technology. The model presented in this paper integrates the management and analysis of information risks and decision-making theory and thus creates a framework for the integrated management information system based on the technological risk assessment and Bayesian learning. The paper shows simulation and two case study scenarios in which is presented a potentially wide range of usage.

Keywords: *Bayesian learning, information system risk, risk assessment, threats, vulnerabilities*

Metodologija odlučivanja temeljena na procjeni rizika informacijskih sustava i Bayesovom učenju

Izvorni znanstveni članak
Procjena rizika je tema kojom se bave kompanije iz širokog spektra djelatnosti i na temelju iste donose važne odluke za buduće poslovanje. Vrlo je važno strateški se opredijeliti i odabrati ključne odluke i unutar sustava upravljanja informacijskim sustavima. Različiti rizici koji proizlaze iz prijetnji i ranjivosti računalne opreme, osoblja koje je zaduženo za upravljanje tom opremom i sustavima za koje je informacijska tehnologija podrška, ugrožavaju temeljni cilj informacijskih sustava, da rade efektivno i efikasno. Procjena rizika informacijskih sustava temelji se na identificiranju prijetnji i ranjivosti, te određivanju vjerojatnosti njihovih ostvarenja, a time i vjerojatnost ostvarenja rizika. U trenutku kada je vjerojatnost događaja opisanog indikatorima koji ga mogu prouzročiti poznata, može se raspravljati i o matematičkim modelima pomoću kojih je moguće izračunati vjerojatnost događaja u budućnosti. Ako je pored procjena, poznata i statistička analiza u obliku zapisa stvarnih događaja, tada je statistički model moguće razviti u ozbiljan alat za podršku odlučivanju prilikom upravljanja informacijskim sustavima. U radu je prikazan model koji objedinjuje procjenu rizika informacijskih sustava i Bayesovu teoriju odlučivanja.

Ključne riječi: *Bayesovo učenje, prijetnje, procjena rizika, ranjivosti, rizici informacijskih sustava*

1 Introduction

An information asset includes all of the physical assets of a company, the staff, but also the processes and activities that can also be analysed in the same way. Decision-making process is based on two segments: mechanical decision is based on the calculation of the probability of favourable or adverse events, and in the case of uncertainty it is possible to seek a decision of man. Machine decision is based on indicators that need to be clearly defined and also scenarios describing edge cases need to be developed. Boundary cases are scenarios when it is not possible to bring a machine decision with a certain probability so that manual decisions are required. Both ways give feedback to the learning process.

Mathematical algorithm in background is underlying the parameters based on probabilities while calculation determines the provided decisions. Risk assessment is included because its parameters are already based on probabilities, and as such explicitly indicate the probabilities of realization of events and the impact of these events on the information systems and business.

Parameters underlying the decision-making process can be different and it is important to determine key performance indicators of effectiveness and efficiency in order to be higher.

2 Risk assessment

Risk analyses can be presented in a format which is almost independent from the application [1]. The most important step in the process of a risk assessment is to identify the context of the decision problem [1], i.e. the relation between the considered engineering system and/or activity and the analyst performing the assessment:

- Who are the decision maker(s) and the parties with interests in the activity (e.g. society, client(s), state and organizations)?
- Which matters might have a negative influence on the impact of the risk assessment and its results?
- What might influence the manner in which the risk assessment is performed (e.g. political, legal, social, financial and cultural)?

Risk is defined as a result of possible impact of threats to exposed vulnerability of information assets. Information assets are presented as values by using the properties of confidentiality, integrity, availability and other properties essential to the organization. The value of information assets is described as the impact level of these properties [2]. Financial value is not practical to use in this case because it is often not easy to determine how valuable information assets in cash are or described with a qualitative assessment. Information assets are necessary to be classified and divided into groups that need to be negotiated in the initial preparation of the risk assessment.

It is convenient to observe and validate information assets in this way, because it is possible to manage wide range of assets. Also it is possible to achieve compliance with other processes in information system management cycle, e.g. Business Continuity Management, Incident Management.

Risk assessment is used in a number of situations with the general intention to indicate that important aspects of uncertainties, probabilities and/or frequencies and consequences have been considered in one way or the other. Decision theory provides a theoretical framework for such analyses, Fig. 1.

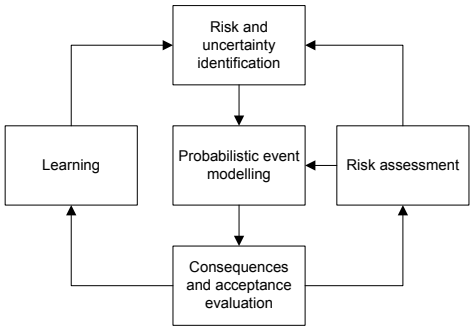

Figure 1 Decision theory based on Risk management

Calculated risks are compared with the accepted risks initially stated in the risk acceptance criteria. If there is no possibility to accept the risks in accordance with the specified risk acceptance criteria, there are principally four different methods to proceed [5]:

- **Implementation of control**: Risk mitigation is implemented by modification of the system such that the source of risk is removed. Risk reduction may be implemented by reduction of the consequences and/or the probability of occurrence – in practice risk reduction is normally performed by a physical modification of the considered system.
- **Risk transfer**: Risk transfer may be performed by e.g. insurance or other financial arrangements where a third party takes over the risk.
- **Avoiding risk**: Avoiding risk may be selected in cases where information resources or the processes are not necessarily required for the proper system operation.
- **Risk acceptance**: If the risks do not comply with the risk acceptance criteria and other approaches for risk treatment are not effective than risk acceptance may be an option.

Risk mitigation methods are representing the results of decision-making process.

Decision support model presented in this article is based on events monitoring and regarding this it needs to be clear how every particular event is processed. Therefore, risk assessment is explained in the paper as a method of analysis of events which results are explicit probabilities.

3 Basic probability rules and Bayesian theorem

In this model, a Bayesian theorem is used and methodology is developed by integrating the database of observed cases with expert experience and knowledge. All monitored parameters are considered trough likelihood of realization of event and thus are completely compatible with the Bayesian thesis [3].

An event E is defined as a subset of the sample space (all possible outcomes of a random quantity) Ω. The failure event E of e.g. a structural element can be modelled by $E=\{R \leq S\}$ where R is the strength and S is the load. The probability of failure is the probability $P_f=P(E)=P\{R \leq S\}$. If a system is modelled by a number of failure events, failure of the system can be defined by a union or an intersection of the single failure events.

a) If failure of one element gives failure of the system, then a union (series system) is used to model the system failure, E:

$$E = \cup_{i=1}^{m} E_i, \tag{1}$$

where E_i is the event that represents failure from i to m number of events.

b) If failures of all elements are needed to obtain failure of the system, then an intersection (parallel system) is used to model the system failure, E:

$$E = \cap_{i=1}^{m} E_i. \tag{2}$$

Disjoint / mutually exclusive events are defined by $E_1 \cap E_2 = 0$ where 0 is the impossible event.
A complementary event E is denoted by $E \cap \bar{E} = 0$ and $E \cup \bar{E} = \Omega$.

The so-called De Morgan's laws related to complementary events are

$$E_1 \cap E_2 = \overline{\bar{E_1} \cup \bar{E_2}} \text{ and } E_1 \cup E_2 = \overline{\bar{E_1} \cap \bar{E_2}}. \tag{3}$$

The conditional probability of an event E_1 given another event E_2 is defined by:

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}. \tag{4}$$

Event E_1 is statistically independent of event E_2 if $P(E_1|E_2) = P(E_1)$.

From (4) we have

$$P(E_1 \cap E_2) = P(E_1|E_2)P(E_2) = P(E_2|E_1)P(E_1). \tag{5}$$

Therefore if E_1 and E_2 are statistically independent $P(E_1 \cap E_2) = P(E_1)P(E_2)$.

Using the multiplication rule in (5) and considering mutually exclusive events E_1, E_2, E_3,..., E_m, the total probability theorem follows:

$$\begin{aligned} P(A) = {} & P(A|E_1)\,P(E_1) + P(A|E_2)\,P(E_2) + \cdots + \\ & + P(A|E_m)\,P(E_m) = \\ & = P(A \cap E_1) + P(A \cap E_2) + \cdots + P(A \cap E_m), \end{aligned} \tag{6}$$

where A is an event.

From the multiplication rule in (5) it follows $P(A \cap E_i) = P(A|E_i)P(E_i) = P(E_i|A)P(A)$.

Using also the total probability theorem in (6) the so-called Bayesian theorem follows from:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{m} P(A|E_j)P(E_j)}. \tag{7}$$

Bayesian theorem allows to determinate the probability of an event based on the probabilities of two or more recorded and independent events [4]. It is

possible to calculate the probability of confirmation of a set of initial hypotheses in case of realized and confirmed event A. In case of the application of this rule, it is necessary to know the probability $P(A)$ and $P(E_i)$, and it is also necessary to know statistical background used to determine the probability $P(A|E_i)$. Formula is valid in the absence of mutual dependence between events A and series of hypotheses and E.

3.1 Assumptions

Probability of an information assets property compromising within a computer system based on statistical data from the past and risk assessments by authorized persons. It is also possible to use objective data published by the authorities [7]. All initial data can be found within the historical database that contains the initial conditions for the operation of the system.

Machine decision is based on the probability calculation of event that belongs to favourable or to adverse events. To be able to make decisions, it is necessary to define the limits of acceptability. In the example described in the paper the boundaries are defined by of the resulting probability of 90 % and all probabilities above this are characterized as unfavourable.

New events and new combinations of parameters which potentially threaten the properties of information assets are recorded by the monitoring systems and stored into the real-time database. New parameters are stored in database and there they were assigned with initial probability values. Decisions on the initial values and the circumstances must be made and regarding this new events are classified.

In the case of advanced mode usage, there is ability to define the level of uncertainty and the resulting range of probabilities. Within events where cannot be assessed enough confidence and thus categorized, it is necessary to determine the limit of probability values. Advanced system also allows additional control of machine decisions of favourable and unfavourable events. Such control must also be mechanical, but it must be based on detailed controls of the system parameters. This section is necessary to be further developed and indicators of suggesting a greater certainty of the correctness or incorrectness of decision need to be defined.

3.2 Specification of previous events probabilities

The historical database contains the probability of previous events. These probabilities need to be calculated on the basis of objective indicators. It is possible to use external sources of knowledge in case of the unavailability of proprietary data. The table shows one part of the database based on which simulation is conducted. The table contains the real data and the number of occurrences of threats and vulnerabilities in the event of information security was initially based on real data. After simulation these numbers have increased in line with the simulated events (Tab. 1).

Table 1 Example of monitoring parameters database

Events parameters	Number of occurrences of the parameter in adverse events	Number of occurrences of the parameter in favourable events	The probability of adverse events	Uncertain events (0,85 ÷ 0,9)
Inadvertent destruction of cables	2	20	0,560000000	
Inadvertent crushing equipment		21	0,377358491	
Damage due to construction works		19	0,401146132	1
Termination of alternative power supply	1	17	0,428134557	1
Voltage fluctuations	1	13	0,494699647	
Termination of internal infrastructure	1	18	0,414201183	

The simulation is conducted by generating randomly selected parameters from the database and adding new parameters. In this way, the artificially generated event can be described by the familiar parameters, but also contains new information.

After the simulation of events the knowledge base contained the following amount of data (Tab. 2).

Table 2 Simulation statistics

Adverse events	133	8,81 %
Favourable events	1377	91,19 %
Number of parameters in database	765	
Uncertain events	7	0,46 %

With every new event there is new mechanical decision based on the calculation of the probability of belonging to a set of favourable or set of adverse events. After deciding, the event is added to the total number of sets in which it is classified. In this way it increases the probability of making a valid machine decision.

3.3 Calculation of the adverse events probability

Adaptations of Bayesian theorem to system and application of the developed algorithm can be represented as:

$$P_{\mathrm{NOK}} = \frac{\frac{n_{\mathrm{NOK}}}{N_{\mathrm{NOK}}}}{\frac{n_{\mathrm{NOK}}}{N_{\mathrm{NOK}}} + \frac{n_{\mathrm{OK}}}{N_{\mathrm{OK}}}},$$

$_{\mathrm{NOK}}$ – Probability of adverse events
n_{NOK} – The number of occurrences in a set of adverse events
n_{OK} – Number of occurrences in a set of favourable events
N_{NOK} – The total number of adverse events
N_{OK} – The total number of favourable events.

Both the relations of occurrence of events and the total number of the same event types are the probabilities of occurrences. In case of insufficient number of events in the database, it is possible to use the probabilities of the information system risk assessment [6].

Risk assessment methodology compatible with the model below must be based on an assessment of probabilities of the realization of threats based on current vulnerabilities. A combination of threats and vulnerabilities makes risks and it is possible to calculate the probability of realization of adverse events, and the consequent likelihood of favourable events.

$$P_{\mathrm{NOK}} = \frac{p_{\mathrm{NOK}}}{p_{\mathrm{NOK}} + p_{\mathrm{OK}}},$$

P_{NOK} – Probability of adverse events
p_{NOK} – Probability of adverse event
p_{OK} – Probability of favourable event.

The use of such a calculation of the probability of adverse events can replace the usage of assumptions in which the parameters of the insufficient number of entries are allocated with initial value of probability.

4 Bayesian learning and decision theory

Bayesian Decision Support System integrates the concept of uncertainty into the risk calculations. This is just a small sampling of the many risk assessment tools available.

In typical decision problems encountered the information basis is often not very precise. In many situations it is necessary to use historical data. The available historical information is often not directly related to the problem considered but to a somewhat similar situation. Furthermore, an important part of a risk assessment is to evaluate the effect of additional information, risk reducing measures and/or changes of the considered problem. It is therefore necessary that the framework for the decision analysis can take these types of information into account and allow decisions to be updated based upon new information. This is possible if the framework of Bayesian decision theory is used [9] and [10].

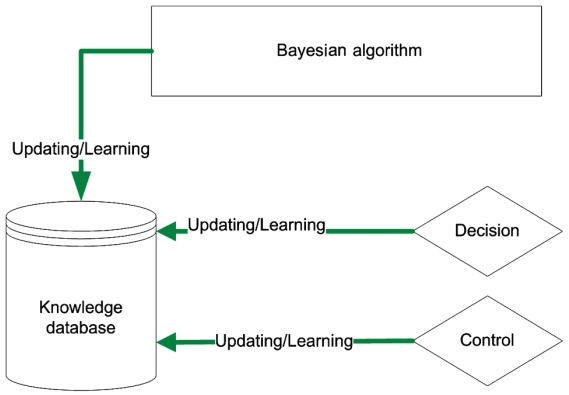

Figure 2 Bayesian learning

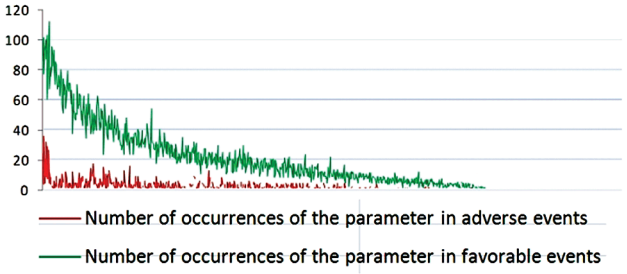

Figure 3 Simulated parameters and machine decisions

A fundamental principle in decision theory is that optimal decisions must be identified as those resulting in the highest expected utility [11]. In typical engineering applications the utility may be related to consequences in terms of costs, fatalities, environmental impact, etc.

In these cases the optimal decisions are those resulting in the lowest expected costs, the lowest expected number of fatalities and so on.

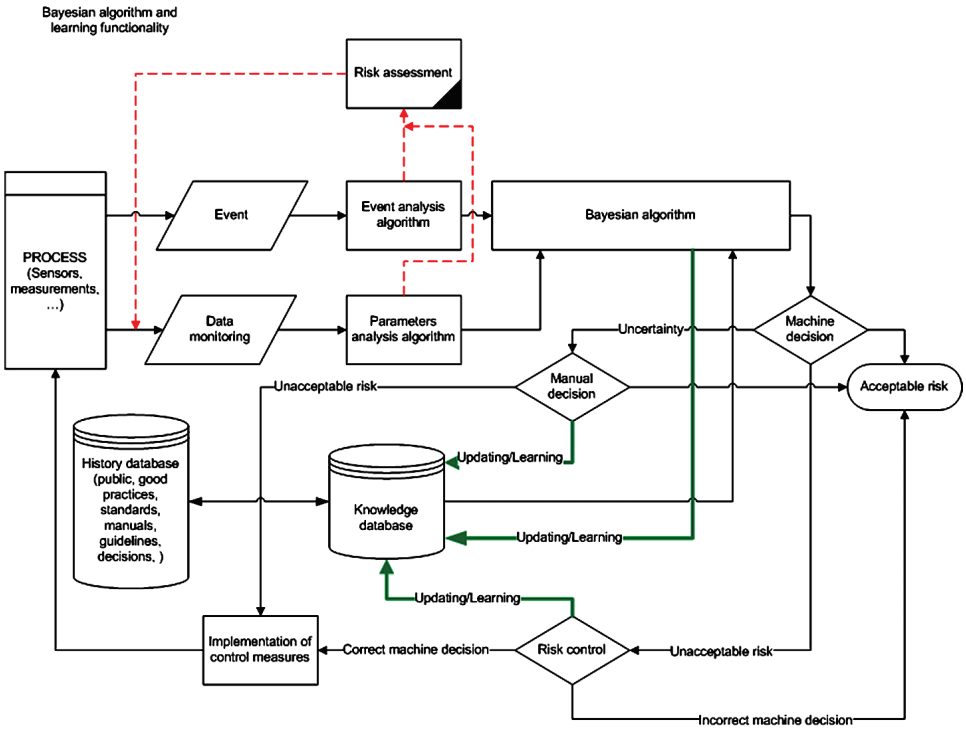

Figure 4 Complete Bayesian decision-making and learning flowchart

Knowledge database can be filled with the data processed by Bayesian algorithm. These data are resulting from real-time monitoring systems that are calculated from Bayesian networks. The greatest amount of new data comes from this part of the system.

Updating the knowledge database filled with the manual decisions occurs only when using a more advanced system that includes the possibility of uncertainty machine decisions.

Additional controls in the case of their existence, creating a link back to the knowledge database and making system up to date and learning functional (Fig. 2).

The diagram shows the amount of parameters assigned to favourable and unfavourable events during the 1500 event simulations randomly generated (Fig. 3).

4.1 Developed model

On the basis of the presented assumptions and theoretical background, a model of decision support is created. Model estimates the probability of threats and vulnerabilities and this implies to the risk of information systems management.

Figure 4 shows an example of this model usage and learning functionality and additional uncertainty range. There is a noted link to risk assessment process which indicates to information between different processes exchanged [8].

4.2 Case study – Information system risk assessment, threats and vulnerabilities

Simulation of decision support systems in the Spam filter case were implemented using some already classified spam messages. A text processing algorithm is developed to analyse incoming messages to separate words and the Bayesian model estimates probability of affiliation of a word to one of the groups: Spam or Ham (expression Ham is used according to common spam filtering syntax [13]). Actual e-mail messages from the Gmail service were used to learn the system (Fig. 5).

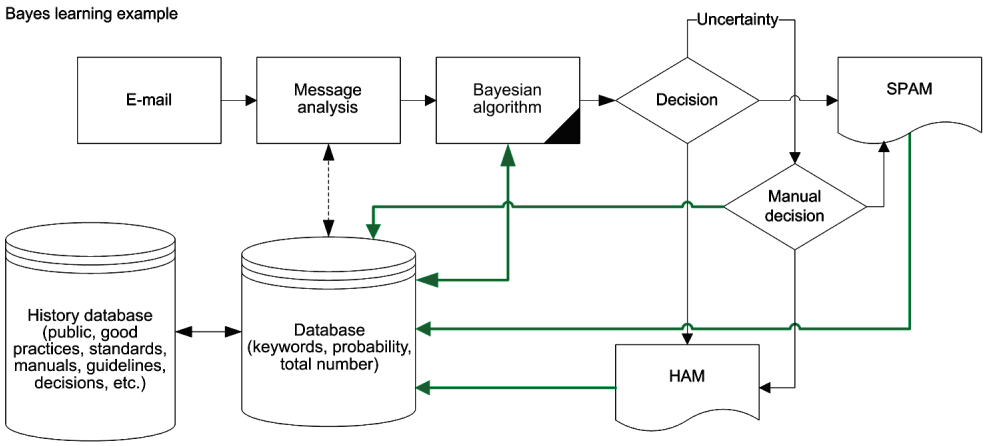

Figure 5 Case study model using in Spam filtering

Learning was based on already classified spam messages from the Trash mailbox and for Ham group from Inbox. Simulation model is selected in a manner in which the threats are spam messages, and the vulnerabilities are caused by improper handling of e-mail system.

Table 3 Sample database in Spam filtering Case study

Token	Spam Frequency	Ham Frequency	Spamicity
bit	4296	2292	0,344723798
blood	383	53	0,669775576
nigerian	140	2	0,862697231
about	3301	2578	0,264373172
account	585	563	0,225789499

Spamicity is calculated probability of unwanted message, which is probability of adverse events, above defined (Tab. 3).

4.3 Results and comparison

Based on 60 messages taken from the e-mail mailbox, a comparison of classification was conducted. 59 messages are equally classified, and one message is in our model marked as Spam, while Google spam filter marked the same message as correct. The message is manually analysed and classified as valid. Error rate in the simulation model is 1,6 % focusing on initial under-developed database. Following a learning period of system increases accuracy and reduces the likelihood of errors.

Some other analyses of Spam solutions showed very similar results. Spam precision is in paper [12] in range 92,3 ÷ 100 %, but with a detailed combination of attributes to mark messages. In this paper, the primary task is not spam filtering, and therefore detailed analysis and additional attributes are not used. The aim is to use Bayesian learning and decision-making in practice case.

A similar experiment was made in case of Intrusion Prevention/Detection System (IPS/IDS) in which the model is compared with Nessus (Tenable Network Security Inc.) and the results also indicate to the compliance of more than 90 %. In this case, there is no analysis of e-mail messages and words as parameters inside, but packets in network traffic. The picture is identical to that shown in Fig. 5, but in the case of IPS/IDS system there is no messages and Spam/Ham decisions, but IP packages and valid/not valid package. Also, there is option to use packages or IP addresses in manner of "black list" definitions.

Table 4 Sample database in Spam filtering Case study

Token	Malicious behaviour	Non-malicious behaviour	Probability of non-malicious behaviour
libvorbis	654	211	0,358941587
openjdk-6	324	26	0,785265877
libxml2	190	5	0,852468874
xulrunner-1.9.2	1980	988	0,325871212
dbus-glib	683	332	0,258749663

Tab. 4 shows some of the characteristic values of this experiment. The parameters (token) in the table are the abbreviation of classified vulnerabilities in the Nessus application. Network of 1523 computers was scanned, servers were not included, and the database is loaded from Nessus. Fig. 6 shows the parameters in the experiment in the form of statistical data for the parameters used.

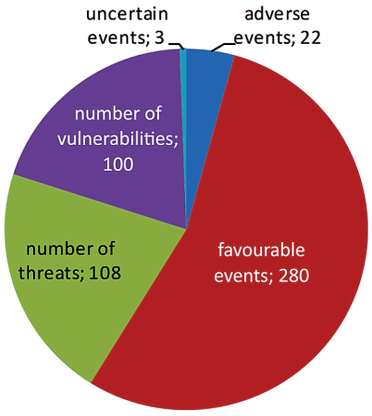

Figure 6 Parameters overview in Intrusion Detection/Prevention case study

Correspondence between classification of Nessus vulnerability and the decisions taken by Bayesian algorithm is 92 %.

5 Conclusion

In this paper, a decision support model of management information system is proposed. Model is based on continuous monitoring of threats and vulnerabilities which make information risks. Information risks managing system has a learning ability based on Bayesian theory.

Simulation shows results compliant with expectations, and it was performed using the actual risk assessment data. Comparison of proposed model results and Google spam filter tool showed significantly better performance and accuracy. Also shown is the comparison with vulnerability testing system Nessus. All results are compared and show compliance in percentage greater than 90 %.

The effectiveness and accuracy of the model are demonstrated through case studies, which indicate that the model is able to improve the accuracy and efficiency of security risk assessment for information systems. The main advantage of this model is its simplicity and flexibility, which make it competitive in the market of large and expensive systems.

Mixture of various applications of the developed algorithm shows a wide range usability and adaptability.

The disadvantages of this model are potentially long period of learning and the need of previous risk assessment data. Data from the risk assessment need to be structured as probabilities. Due to the large differences in risk assessment approaches we recommend to use the described methodology or similar rating of used parameters.

Future work will focus on applying the proposed model to other practice situations, and building more sophisticated constraints into the model to enhance the performance of managed information systems.

6 References

[1] Avena, T.; Zio, E. Some considerations on the treatment of uncertainties in risk assessment for practical decision making. // Reliability Engineering & System Safety. 96, 1(2011), pp. 64-74.

[2] Landoll, D. J. The security risk assessment handbook, Auerbach Publications, 2006.

[3] Robert, C. P. The Bayesian Choice, From Decision-Theoretic Foundations to Computational Implementation, Springer, 2007.

[4] Kjaerulff, U. B.; Madsen, A. L. Bayesian networks and influence diagrams: A guide to construction and analysis, New York, Springer Science, 2013

[5] Sørensen, J. D. Structural Reliability Theory and Risk Analysis, Aalborg, 2004.

[6] Jensen, F. V. Bayesian networks and decision graphs, Springer-Verlag, New York, 2001.

[7] Pollino, C. A.; Hart, B. T. Developing Bayesian network models within a Risk Assessment framework, International Congress on Environmental Modelling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making, 2008.

[8] Yang, X.; Luo, H.; Fan, C.; Chen, M.; Zhou, S. Analysis of risk evaluation techniques on information system security. // J. Comput. Appl. 28, 8(2008), pp. 1920-1924.

[9] Hjort, N. L.; Nonparametrics, B. Cambridge Series in Statistical and probabilistic Mathematics, Cambridge University Press, 12. 4. 2010.

[10] Devore, J. L. Probability and Statistics for Engineering and the Sciences, Brooks/Cole, Cengage Learning, 2012.

[11] Fisel, J.; Berkes, P.; Orbán, G.; Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. // Trends in Cognitive Sciences, Elsevier. 14, 3(2010), pp. 119-130.

[12] Androutsopoulos, I.; Koutsias, J.; Chandrinos, K. V.; Paliouras, G.; Spyropoulos, C. D. An Evaluation of Naive Bayesian Anti-Spam Filtering, Software and Knowledge Engineering Laboratory, Greece, 2010.

[13] Ghaoui, C. Encyclopedia of human Computer Interaction, Liverpool John Mores University, UK, 2006.

Authors' addresses

Hrvoje Očevčić, Mr. Sc.
Hypo Alpe-Adria-Bank d.d.
Slavonska avenija 6, 10000 Zagreb, Croatia
E-mail: hrvoje.ocevcic@hypo-alpe-adria.hr

Krešimir Šolić, BSc
Medicinski fakultet Osijek
Josipa Huttlera 4, 31000 Osijek, Croatia
E-mail: kresimir@mefos.hr

Krešimir Nenadić, doc. dr. sc.
Elektrotehnički fakultet Osijek
Cara Hadrijana bb, 31000 Osijek, Croatia
E-mail: kresimir.nenadic@etfos.hr