

УДК 519.7

## АНАЛІЗ БАЗИ ДАНИХ ІНФОРМАЦІЙНИХ ЗАГРОЗ МЕТОДАМИ ТЕХНОЛОГІЇ DATA MINING

Л. Білий<sup>1</sup>, С. Копитко<sup>2</sup>

<sup>1</sup> Інститут Банківської справи УБС НБУ, д.т.н., професор  
79000 м. Львів, проспект Шевченка 9

<sup>2</sup> Університет банківської справи НБУ(м. Київ), аспірант  
79000 м. Львів, проспект Шевченка 9  
E-mail: [aspirantura@lbi.wubn.net](mailto:aspirantura@lbi.wubn.net)

Розглядається проблема удосконалення методики проектування ефективних систем захисту інформації з використанням методів data mining. Запропонована модель пошуку прихованих залежностей у формі if-then-правил між характеристиками інформаційних загроз на підставі аналізу записів табличної БД. З допомогою системи інтелектуального аналізу даних WizWhu отримано сукупність правил, які пов'язують деякий набір характеристик інформаційних загроз з очікуваними втратами від атак такого типу на комп'ютерну систему. Здійснено аналіз змісту правил та виявлено на їх підставі переважні впливи деяких залежних чинників на обсяги втрат при здійсненні атаки на комп'ютерну інформацію.

Ключові слова: захист інформації, проектування систем захисту, ефективність систем захисту, data mining, інформаційні загрози, БД інформаційних загроз, if-then-правила, система WizWhu

Сфера застосування технології Data Mining нічим не обмежена, єдиною вимогою є лише наявність необхідної для аналізу вхідної статистичної інформації. Сьогодні, методи Data Mining широко використовують комерційні підприємства, досвід яких показує, що віддача від застосування даної технології може досягати 1000%. Наприклад, відомі повідомлення про економічний ефект, який в 10 – 70 разів перевищив початкові затрати від \$350 до \$750 тис.[1, с.16]. Також відома інформація про проект в розмірі \$20 млн, який окупив себе всього за 4-и місяці. Оскільки головною метою нашої наукової роботи є оцінка втрат від несанкціонованого доступу до комп'ютерної інформації, ми вважаємо доцільним використати переваги технології Data Mining для більш глибокого і якісного аналізу поставленої задачі.

Проблема проектування ефективної системи захисту комп'ютерної інформації обов'язково враховує характеристики множини можливих інформаційних загроз. Оскільки опис загрози складається з деякої досить великої множини характеристик, то досить складно виявити наявні між ними взаємозв'язки. Причому ряд таких зв'язків є неочевидними або прихованими і їх можна діагностувати тільки шляхом аналізу масиву статистичних даних атак на комп'ютерні системи. Технологія Data Mining призначена для виявлення прихованих залежностей між факторами шляхом аналізу історичних масивів даних. Більше того, іншими методами такі залежності не виявляються [2, с 14-16].

Нами у [3, 4] розглядається задача класифікації множини інформаційних загроз з допомогою нейромережевих методів з метою виділення типових загроз, узагальнені характеристики яких передбачається враховувати при розробці ефективних систем захисту комп'ютерної інформації. Для цього нами було створено на підставі описів реальних атак [5]

на інформаційні системи табличну БД інформаційних загроз в середовищі MS Excel. Цікаво було би проаналізувати її з допомогою технології Data Mining.

### 1. Постановка задачі.

Метою цього наукового дослідження було виявлення прихованих залежностей між характеристиками реальних загроз комп'ютерній інформації, описи яких наявні у табличній БД [3]. Для реалізації поставленої мети потрібно було вирішити такі завдання:

- формалізувати задачу аналізу табличної БД з допомогою технології Data Mining;
- вибрати форму представлення прихованих залежностей між характеристиками інформаційних загроз;
- вибрати систему інтелектуального аналізу даних, з допомогою якої власне і буде здійснюватись аналіз табличної БД інформаційних загроз;
- здійснити аналіз табличної БД вибраною системою Data Mining;
- проаналізувати отримані набори прихованих залежностей між характеристиками інформаційних загроз.

Дальше описуються основні отримані результати.

### 2. Формалізація задачі аналізу табличної БД методами Data Mining.

Подібна задача аналізу БД моніторингу точності прогнозів соціально-економічних процесів розглядається у [6], де був запропонований спосіб її формалізованого опису. Так як наша проблема схожа з описаною у [6], то адаптуємо цю модель до нашої БД інформаційних загроз.

Отже, аналогічно як у [6], позначимо через  $B(X(t), D(t), Z(t))$  табличну БД інформаційних загроз станом на момент часу  $t$ . Як бачимо, структурно БД складається з трьох компонент, а саме:

- набору  $X(t)$  характеристик інформаційних загроз, де множина  $X(t) = \{x_1, x_2, \dots, x_n\} \cup l$ . (1)

Через  $x_1, x_2, \dots, x_n$  позначено окремі незалежні характеристики загрози комп'ютерній інформації, а  $l$  – цільова (залежна) характеристика, в якості якої виступає чинник “очікувані втрати від несанкціонованого доступу”;

- множина доменів  $D(t)$ , які визначають на момент часу  $t$  допустимі значення характеристик загроз:  $D(t) = \{D_1(t), D_2(t), \dots, D_n(t)\}, x_i \in D_i(t)$ ; (2)

- сукупності записів  $z_j \in Z(t)$ , причому запис  $z_j$  характеризує  $j$ -ту загрозу  $z_j = \langle x_{j1}, x_{j2}, \dots, x_{ji}, \dots, x_{jn}, l_j \rangle$ . (3)

Зауважимо, що зазначені компоненти можуть змінюватись у часі.

Нехай  $Y \subseteq X(t) \setminus l$  – деяка підмножина набору характеристик  $X(t)$  табличної БД інформаційних загроз. Позначимо через  $\Psi$  множину інтервалів  $[\alpha_k; \beta_k) = \psi_k$ ,  $\psi_k \in D_l(t)$ , де  $D_l(t)$  означає допустиму множину значень цільової характеристики  $l$  на момент часу  $t$ . Ясно, що у деяких випадках  $\alpha_k = \beta_k$  і тоді інтервал вироджується у точку, тобто значення характеристики  $l$  приймає фіксоване значення із  $D_l(t)$ . Тоді формально задачу пошуку прихованих залежностей між характеристиками інформаційних загроз можна сформулювати таким чином: для заданого набору незалежних характеристик  $Y \subseteq X(t) \setminus l$  і множини  $\Psi$  інтервалів зміни залежної змінної  $l \in X(t)$  на підставі сукупності записів  $Z(t)$  табличної БД інформаційних загроз  $B(X(t), D(t), Z(t))$  визначити такі залежності  $l = f_{kr}^t(\gamma, Y, \Psi)$  та  $l = \overline{f_{kr}^t}(\gamma, Y, \Psi)$ , які би задовольняли вимоги

$$P\{f_{kr}^t(\gamma, Y, \Psi) \in \psi_k : z_j \in B(X(t), D(t), Z(t))\} \geq \gamma; \quad (4)$$

$$P\{\overline{f_{kr}^t}(\gamma, Y, \Psi) \notin \psi_k : z_j \in B(X(t), D(t), Z(t))\} \geq \gamma, \quad (5)$$

де  $\gamma$  – рівень достовірності залежності, а  $P\{A\}$  означає ймовірність настання події  $A$ . Залежність (4) визначає позитивний факт, а (5) є заперечною.

Таким чином, модель (1) – (5) дозволяє визначити на момент часу  $t$  для вхідних параметрів  $Y$ ,  $\Psi$  та  $\gamma$  існуючі у БД  $B(x(t), D(t), Z(t))$  залежності виду (4), (5) обсягів втрат  $l$  від обставин приведення інформаційних атак, що описуються характеристиками підмножини  $Y$ . На практиці традиційно використовують дві форми зображення прихованих залежностей [3]: правила *if – then* або *дерева рішень*. Для подальшої конкретизації форми представлення прихованих залежностей (4), (5) вибрали правила *if – then*. Тоді можна конкретизувати вигляд  $f_{kr}^t(\gamma, Y, \Psi)$  та  $\bar{f}(\gamma, Y, \Psi)$  для усіх  $\psi_k \in \Psi$ :

$$f_{kr}^t(\gamma, Y, \Psi): \text{if}(x_{i_1} = d_{i_1}) \wedge ((x_{i_2} = d_{i_{21}})) \wedge \dots \wedge (x_{i_m} = d_{i_m})$$

$$\text{then } l \in \psi_k, x_{i_{mi}} \in Y, d_{i_m} \in D_{i_m}; \quad (6)$$

$$\bar{f}_{kr}^t(\gamma, Y, \Psi): \text{if}(x_{i_1} = d_{i_1}) \wedge ((x_{i_2} = d_{i_{21}})) \wedge \dots \wedge (x_{i_m} = d_{i_m})$$

$$\text{then } l \notin \psi_k, x_{i_{mi}} \in Y, d_{i_m} \in D_{i_m}. \quad (7)$$

Саме у формі правил *if – then* виду (6), (7) будемо шукати приховані залежності у нашій табличній БД інформаційних загроз.

### 3. Загальний опис табличної БД інформаційних загроз.

Запропонована нами задача класифікації множини інформаційних загроз [3, 4] передбачає розбиття кожної загрози на низку незалежних характеристик, які виступають вхідними параметрами нашої моделі. Саме тому зібрану статистичну інформацію [5] було адаптовано до нашої проблеми і створено табличну БД необхідного формату. Загальна характеристика БД інформаційних загроз представлена у табл. 1.

### 4. Апробація моделі аналізу табличної БД інформаційних загроз.

Достатньо велика кількість характеристик потребує додаткових методів аналізу їх взаємозалежності. Для вирішення цієї проблеми застосовано програмний пакет *WizWhu* [3], який дозволяє знайти приховані залежності у створеній нами БД інформаційних загроз.

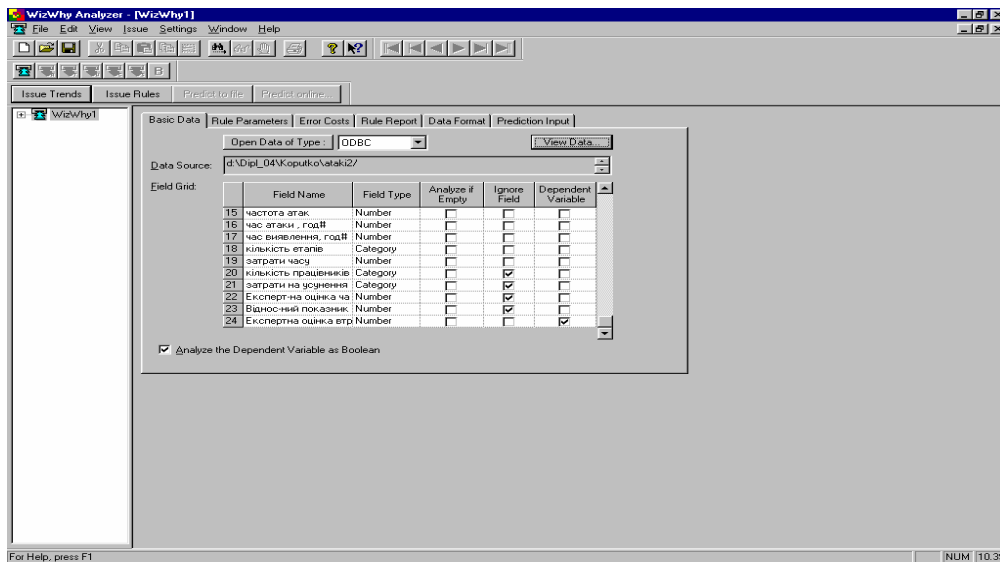
Таблиця 1

Характеристика БД інформаційних загроз

Назва характеристики	Всього	З них описують				
		спрямованість загрози	систему захисту	обставини атаки	обставини виявлення атаки	оцінка результату загрози
Кількість факторів в описі загрози	21	5	7	4	3	2
Кількість якісних характеристик	14	5	7	1	1	0
Кількість записів у БД	15					

Джерело: [5; власні розрахунки]

Процес автоматизації пошуку прихованих залежностей включає етапи визначення джерела даних для аналізу, вибір множини незалежних факторів  $Y$  із сукупності характеристик загроз, введення параметрів пошуку правил та активізації власне процесу відшукування правил *if-then*. Зразок діалогової панелі для реалізації перших двох етапів показано на рис. 1.

Рис. 1. Панель для вибору джерела даних та факторів множини  $Y$ .

Вхідні параметри для системи WizWhu, вибрані для пошуку правил, наведені у табл. 2. Зазначимо, що для спрощення використали лише два інтервали зміни значення вірогідних обсягів втрат від інформаційної загрози. Уведення параметрів пошуку здійснювали через вікно, показане на рис. 2.

Таблиця 2

Параметри системи WizWhu для пошуку прихованих залежностей у БД інформаційних загроз

п/п	Назва параметра	Значення
	Розбиття інтервалу зміни обсягів втрат	$\Psi = \{\psi_1, \psi_2\}$ , $\psi_1 = (30;60], \psi_2 = (60;80]$
	Мінімальна вірогідність правила if - then	80%
	Мінімальна вірогідність правила if-then-Not	80%
	Максимальна кількість умов	4
	Залежний фактор	Обсяги втрат
	Кількість незалежних факторів	18

Джерело [власні розрахунки].

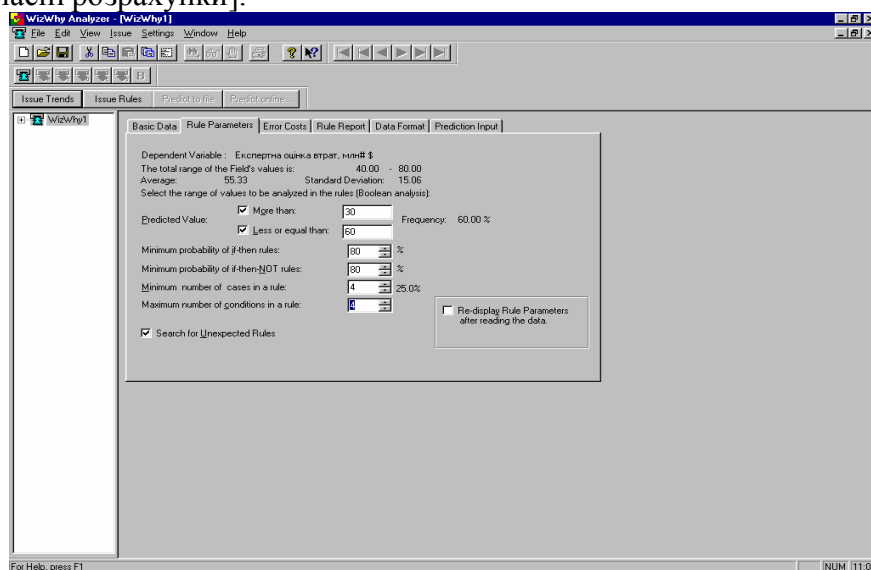


Рис. 2. Панель уведення параметрів пошуку правил системою WizWhu.

Активізувавши процес аналізу записів табличної БД інформаційних загроз, отримали деяку сукупність правил виду (6), (7), які система видає у формі трьох звітів, серед яких найважливішими є звіти про віднайдені правила і так звані *неочікувані правила*. Змістовна інтерпретація знайдених if-then-правил для двох інтервалів зміни цільового чинника включена у табл. 3

Таблиця 3

Узагальнена характеристика залежностей у БД інформаційних загроз для системи інтервалів  $\Psi = \{\psi_1, \psi_2\}$  зміни очікуваних обсягів збитків.

Інтервал зміни втрат	Кількість правил		Кількість факторів	Зміст правил
	всього	заперечних		
30 – 60 млн. \$	14	–	11	1) для невеликої організації з ОС Windows 2000 слід очікувати обсягів втрат з інтервалу 30 – 60 млн. \$ від кожної атаки; 2) атака на комп'ютерну систему з компонентами захисту і Web-сервером без спеціального програмного забезпечення, що здійснюється раз на період, зумовлює у 85,7% випадків обсяги втрат у 30–60 млн. \$; 3) у 80% випадків атаки на сервери невідомого типу або на систему у довільний період доби, або на інформаційні системи невеликих організацій чи без чітко окресленої мети і з частотою раз на період та загрози, що здійснюються у два етапи, характеризуються обсягами збитків, які попадають у інтервал 30–60 млн. \$; 4) атаки на Web-сервер без спеціального програмного забезпечення у будь-який період доби дають втрати з інтервалу 30–60 млн. \$;
60 – 80 млн. \$	14	14	11	Системою отримано набір if-then-правил, у яких умовна частина (фрагмент між if та then) співпадає з відповідним правилом для інтервалу 30–60 млн. \$, а заключна частина (після then) є запереченням належності значення залежного фактора до інтервалу 60–80 млн. \$. Тому узагальнений зміст цієї сукупності правил можна отримати із попередніх шляхом заперечення висновку у правилі щодо інтервалу 60–80 млн. \$.

Джерело: [звіти програми *WizWhu*; власні розрахунки].

Цікавими є висновки системи щодо важливості характеристик опису інформаційної загрози у відшуканих правилах. Ця інформація включена у табл. 4. Як бачимо, із 18 незалежних факторів, що формували нашу множину  $Y$ , лише 11 були використані у правилах для обох інтервалів зміни цільового чинника. З них 4-й, 7-й і 9-й характеризують обставини проведення атаки, 1-3-й і 6-й та 10-й визначають характеристики системи захисту інформації, 5-й - об'єкт загрози, а 11-й оцінює часові затрати на усунення збоїв у системі.

Також заслуговує на увагу ранжування факторів за значущістю, тобто за частотою їх появи у правилах. На першому місці знаходиться чинник якості програмного забезпечення ІС, на другому – вид сервера, а третє місце посідають декілька факторів (операційна система як базовий елемент захисту комп'ютерної системи, час здійснення атаки, мета загрози). Тому при розробці ефективних систем захисту інформації слід звертати увагу в першу чергу саме на ці чинники.

Таблиця 4

Характеристика незалежних факторів у віднайдених системою WizWhu if-then правил

№ п/п	Назва фактора	У правилах для інтервалу			
		використання		Вага фактора, %	
		$\Psi_1$	$\Psi_2$	$\Psi_1$	$\Psi_2$
1	Операційна система	+	+	21,43	21,43
2	Програмне забезпечення	+	+	28,57	28,57
3	Компоненти захисту	+	+	14,29	14,29
4	Частота атак	+	+	14,29	14,29
5	Організація	+	+	14,29	14,29
6	Вид сервера	+	+	24,43	24,43
7	Період доби	+	+	21,43	21,43
8	Кількість етапів	+	+	7,14	7,14
9	Мета загрози	+	+	21,43	21,43
10	Утиліти	+	+	7,14	7,14
11	Час виявлення	+	+	7,14	7,14

Джерело: [звіти програми WizWhu; власні розрахунки.

Висновки.

Перш за все відзначимо доцільність застосування технології data mining для аналізу інформаційних масивів зі статистикою атак на комп'ютерну інформацію. Як свідчить апробація нашої моделі на масиві описів реальних інформаційних загроз, можна очікувати виявлення залежностей очікуваних втрат від атаки на комп'ютерну систему залежно з обставинами як проведення самої атаки, так і характеристик інформаційної системи та системи захисту. Такі залежності, очевидно, потрібно враховувати у процесі проектування ефективних систем захисту комп'ютерної інформації.

Звернемо увагу також на той факт, що система WizWhu для другого інтервалу зміни очікуваних обсягів втрат знайшла лише заперечні правила, умовна частина яких співпадає з відповідними правилами першого інтервалу. Причиною цього є тільки обмежений поки-що обсяг табличної БД інформаційних загроз. Тому подальшим нашим завданням буде поповнення нашої табличної БД новими описами інформаційних загроз з метою аналізу її методами технології data mining.

1. Кречетов Н. Продукты для интеллектуального анализа данных// Рынок программных средств. - 1997. - № 14-15. - С. 32-39.
2. Дюк В., Самойленко.А. Data mining: учебный курс. – СПб: Питер, 2001. – 368 с.
3. Копитко С.Б. Класифікація загроз комп'ютерній інформації з використанням карт Кохонена//Вісник Львів. ун-ту. Серія екон. – 2007. - Вип.37(1). - С.586-591.
4. Копитко С.Б. Комплекс економіко-математичних моделей оцінювання ефективності захисту комп'ютерної інформації.//Науково-виробничий журнал “Держава та регіони”. Серія: Економіка та підприємництво. – 2007. - №6. - С. 58 – 63.
5. Ученик по защите от хакеров//CD “Эксперт: криптографическая защита данных”. – 2004, Новые технологии.

6. Твердохліб І.П. Удосконалення методології економіко-математичного прогнозування на підставі методів data mining. // Праці III –ї міжнародної школи-семінару “Теорія прийняття рішень” (м. Ужгород, 2-7 жовтня 2006.). – Ужгород: УжНУ, 2006. - С.84-85.

## ANALYSIS OF BASE OF THESE INFORMATIVE THREATS BY METHODS OF TECHNOLOGY OF DATA MINING.

L. Bilyj<sup>1</sup>, Ń. Kopytko<sup>2</sup>

<sup>1</sup> Institute of banking of UBS of NBU, 79000 city Lvov, boulevard of Shevchenko 9

<sup>2</sup> Universities of banking of NBU(i. Kyiv), 79000 city Lvov, boulevard of Shevchenko 9

E-mail: [aspirantura@lbi.wubn.net](mailto:aspirantura@lbi.wubn.net)

*The problem of improvement of design technique effective systems of defence of information is examined with the use of methods of data mining. The model of search of the hidden dependences is offered in form if-then-правила between descriptions of informative threats on the basis of analysis of records of tabular DB. With the help of the system of intellectual analysis of information of WizWhu the aggregate of rules which bind some set of descriptions of in of structure threats to the expected losses from the attacks of such to the type on the computer system is got. The analysis of maintenance of rules is carried out and found out on their foundation overwhelming influences of some dependent factors on the volumes of losses during realization of attack on computer information.*

*Keywords: defence of information, planning of the systems of defence, efficiency of the systems of defence, data mining, informative threats, DB of informative threats, if-then-правила, system of WizWhu*