ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ЗАХИСТУ ІНФОРМАЦІЇ В ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ СИСТЕМАХ

ЛЕКТОР: професор кафедри
безпеки інформаційних і комунікаційних систем
д.т.н., с.н.с. Грищук Руслан Валентинович.

МОДУЛЬ 2

МОДЕЛІ, МЕТОДИ ТА ПІДХОДИ ДО ПОБУДОВИ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ЗАХИСТУ ІНФОРМАЦІЇ

ТЕМА 5
«Методологія побудови ІСЗІ в ІКС»

ЛЕКЦІЯ 6
Методи та моделі інтелектуальних обчислень в системах захисту інформації.

Література:

1. Васильев В. И. Интеллектуальные системы защиты информации / В. И. Васильев. – М. : Машиностроение, 2013. – 172 с.

2. Башмаков А.И. Интеллектуальные информационные технологии: Учебн. пособие / А.И. Башмаков, И.А. Башмаков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. – 304.

Навчальні питання:

1. Стан проблеми та нові рішення.

2. Види моделей інтелектуальних обчислень.

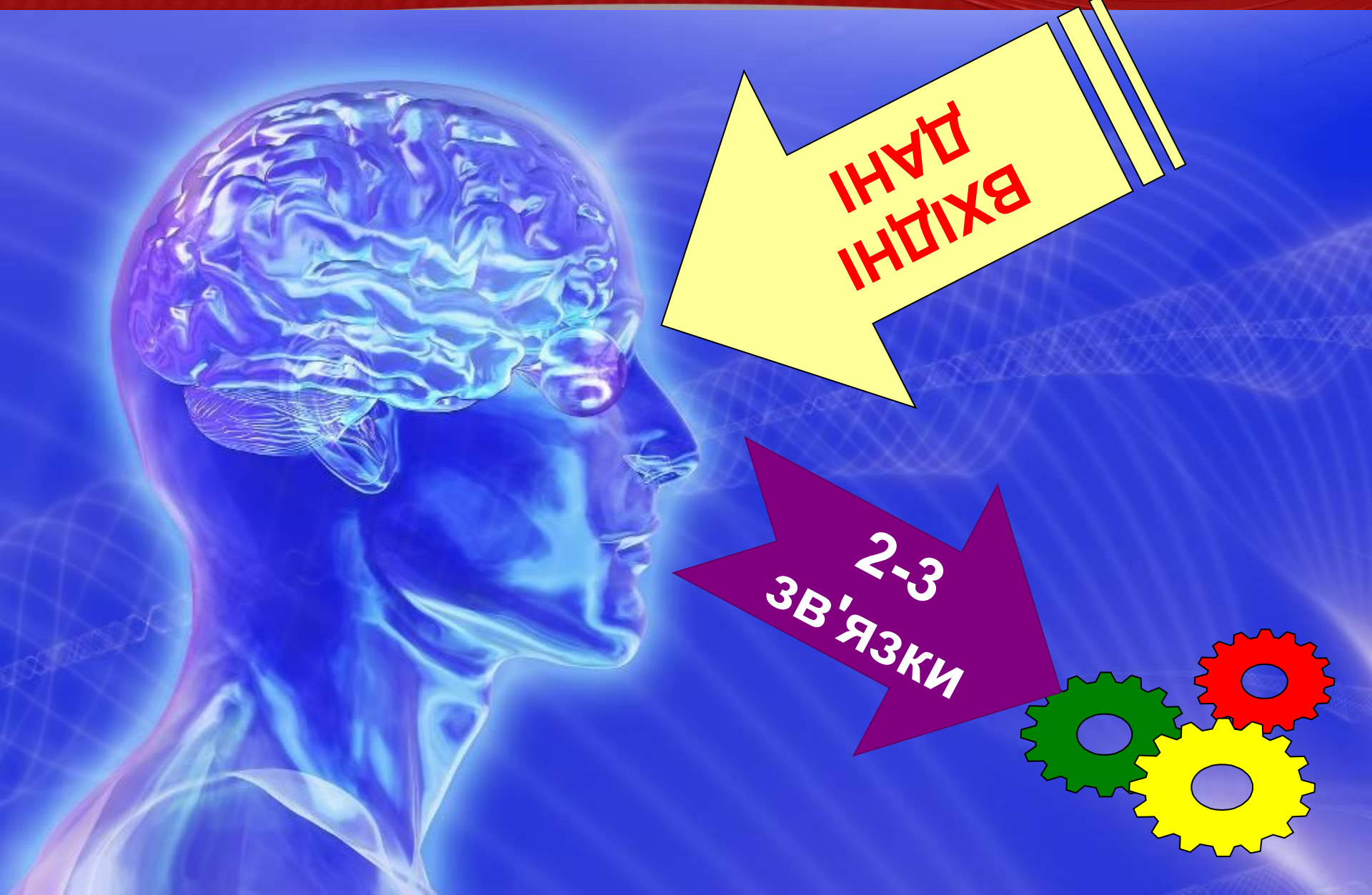
2. Методи інтелектуальних обчислень.

1. Стан проблеми та нові рішення.

Сучасні підходи до пошуку прихованих зв'язків

Застосування інтелектуальних обчислень в системах захисту інформації має на меті суттєве підвищення ефективності їх роботи за рахунок знаходження прихованих правил та закономірностей у мережевих з'єднаннях.

Проблема пошуку прихованих зв'язків

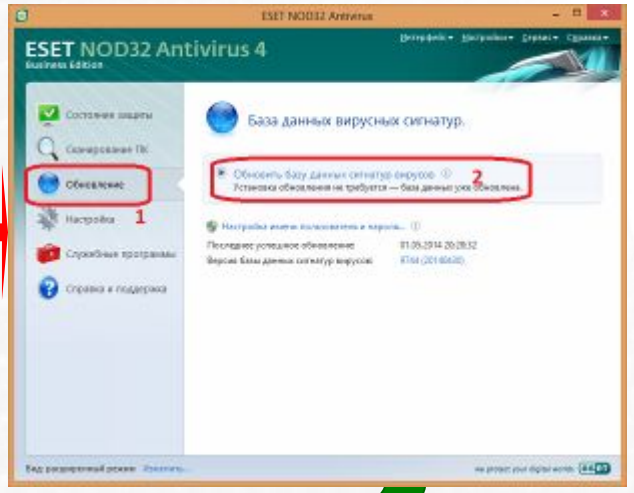


Сучасний стан проблеми

запит

відповідь

Мер
еже
Ве
з'єд
нан
ня

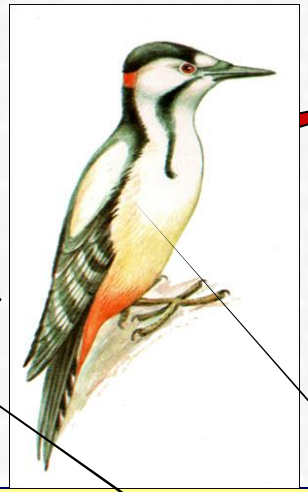


Нові рішення

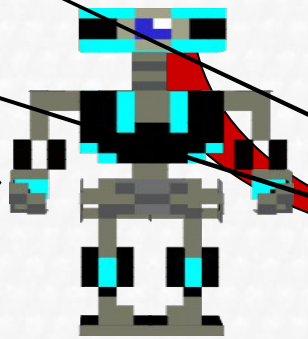
Знаходження прихованих правил та закономірностей у мережевих з'єднаннях на сьогодні може здійснюватися на основі сучасних технологій інтелектуального аналізу даних.

Технології інтелектуального аналізу даних в автоматичному режимі забезпечують пошук шаблонів нормальної поведінки системи та шаблонів атак.

Можливий підхід до організації інтелектуального аналізу даних



Дятел



Aybo

Математична статистика $\frac{1}{n}\sum_{i=1}^{n} S_i$

СИГНАТУРИ

База даних сигнатур ШПЗ

клієнти

ШПЗ

Таким чином,

інтелектуальний аналіз даних в системах захисту інформації – це процес виявлення з накопичуваних в результаті оброблення мережевих з'єднань наперед невідомих закономірностей для побудови шаблонів поведінки ІКС.

Отже,

необхідним атрибутом технології інтелектуального аналізу даних системами захисту інформації в ІКС є клієнт-серверна архітектура системи.

Це забезпечує виконання найбільш трудомістких процедур обробки потоків вхідних даних на високопродуктивному сервері як комп'ютерним вірусологам (“дятлам”), так і роботам (Aybo).

2. Види моделей інтелектуальних обчислень.

Модель - система математичних співвідношень, які описують досліджуваний процес або явище.

Види моделей інтелектуальних обчислень

1. Вид моделей: прогнозуючі моделі ІАД

- класифікаційні моделі;
- регресійні моделі;
- моделі прогнозування часових послідовностей (рядів).

2. Вид моделей: закономірно описові моделі ІАД

- кластеризаційні моделі;
- моделі асоціацій;
- моделі послідовностей.

2.1. Прогнозуючі моделі ІАД в системах захисту інформації ІКС.

Класифікація –

найпоширеніша модель інтелектуального аналізу даних в СЗІ ІКС.

Забезпечує класифікацію тих шаблонів поведінки ІКС, які належать до вже відомих типів кібератак.

Дана функція реалізується тільки за умови аналізу вже класифікованих кібератак та побудованих на їх основі правил класифікації.

Дефініції

МЕТОД КЛАСИФІКАЦІЇ – це правила створення системи класифікаційних угруповань та визначення зв'язків між ними.

ієрархічний

фасетний

СИСТЕМА КЛАСИФІКАЦІЇ – це сукупність методів і правил розподілу множини об'єктів на підмножини відповідно до ознак схожості або несхожості.

ОБ'ЄКТ КЛАСИФІКАЦІЇ – елемент класифікаційної множини.

КЛАСИФІКАЦІЙНЕ УГРУПУВАННЯ – підмножина об'єктів, отриманих у результаті класифікації.

Ієрархічний метод класифікації
(послідовний поділ множини об'єктів на підлеглі класифікаційні угрупування)

ножина вхідних об'єктів
послідовно поділяється
на багаторівневі угрупування.

Ієрархічний метод класифікації

Характеристики:

- ***глибинна*** *(кількість рівнів класифікації);*
- ***ємність*** *(глибина та кількість створених на кожному рівні угрупувань);*
- ***гнучкістю*** *(пристосований для ручної обробки та великих масивів даних).*

Недоліки:

– жорсткість структури;

– неможливість виявлення об'єктів у разі довільного поєднання ознак.

Переваги: *– логічна побудова класифікатора;*

– чітке визначення класифікаційних ознак;

– робота з великими масивами даних;

– зручний у використанні.

Фасетний метод класифікації
(паралельний поділ множини об'єктів на незалежні класифікаційні угруповання)

Фасета – набір значень однієї ознаки класифікації.

Принцип: система класифікації подається у вигляді переліку незалежних фасетів, які містять значення ознак класифікації.

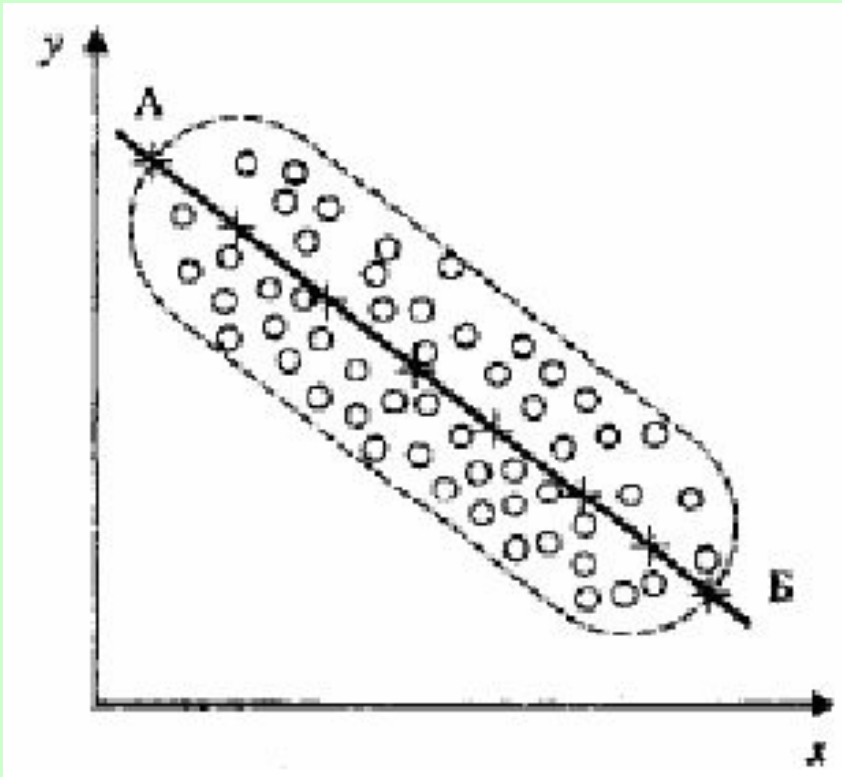
	ФАСЕТИ		
Назва ознаки	Обличчя	Долоня	Автомобіль
Значення ознаки	Чоловіче	Ліва	Легковий
	Жіноче	права	Вантажний
	Значення фасет		

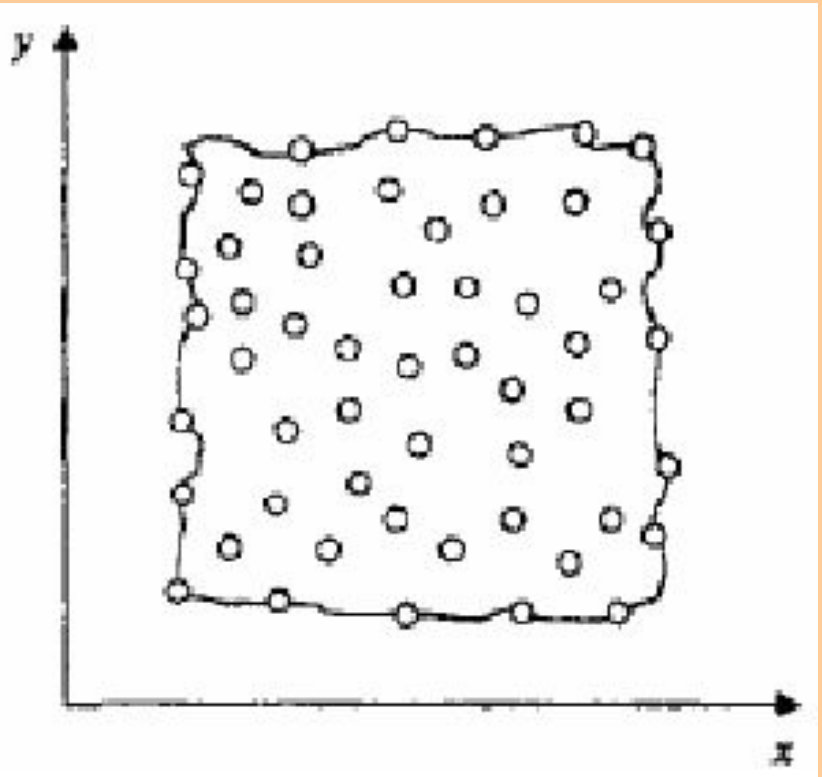
Недоліки: ускладнюється робота з великими масивами даних.

Переваги: гнучка структура класифікації.

РЕГРЕСІЙНИЙ АНАЛІЗ ДАНИХ
(дослідження закономірностей зв'язку між даними)

СУТЬ : встановлення рівняння регресії. → **ОДНОФАКТОРНЕ**, **ДВОФАКТОРНЕ**





РЕГРЕСІЙНИЙ АНАЛІЗ ДАНИХ

постановка задачі

$X = \{x_1, x_2, \ldots, x_i\}$ – множина вхідних даних, $i = \overline{1,n}$

(n - загальна кількість);

$Y = \{y_1, y_2, \ldots, y_i\}$ – множина даних з пам'яті еталонів.

Потрібно знайти регресійну модель $f(x) = ?$

РЕГРЕСІЙНИЙ АНАЛІЗ ДАНИХ

розв'язок задачі в загальному вигляді

Згідно з МНК критерій має вигляд

$$[y_i - f(x_i)]^2 \to min\,. \qquad (1)$$

Нехай $f(x)$ **описується лінійною залежністю вигляду**

$$f(x) = a + bx\,, \qquad (2)$$

де a **та** b **- коефіцієнти, що підлягають визначенню.**

РЕГРЕСІЙНИЙ АНАЛІЗ ДАНИХ

розв'язок задачі в загальному вигляді

Складемо квадратичну форму φ

$$\varphi = \sum_{i=1}^{n} (y_i - a - bx_i)^2, \qquad (3)$$

Необхідні умови:

$$\begin{cases} \dfrac{\partial \varphi}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0; \\ \dfrac{\partial \varphi}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0. \end{cases}$$

Розв'язок СЛАР

$$\begin{cases} a = \dfrac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}; \\ b = \dfrac{\sum_i y_i - a\sum_i x_i}{n}. \end{cases}$$

Оцінка ступеня відхилення

$$\begin{cases} S_b = \sqrt{\dfrac{\sum_i (y_i - bx_i - a)^2}{(n-2)\sum_i \left(x_i - \dfrac{\sum_i x_i}{n}\right)^2}}; \qquad (4) \\ S_a = \sqrt{\dfrac{\sum_i (y_i - bx_i - a)^2}{(n-2)} \left(\dfrac{1}{n} + \dfrac{\left(\dfrac{\sum_i x_i}{n}\right)^2}{\sum_{i=1}^{n} \left(x_i - \dfrac{\sum_i x_i}{n}\right)^2} \right)} \qquad (5) \end{cases} \qquad (6)$$

2.2. Закономірно описові моделі ІАД в СЗІ ІКС.

ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(передбачення поведінки системи в майбутньому)

СУТНІСТЬ *– історична інформація (вікна попереднього спостереження), що є особливими ознаками часу:* ***ієрархія періодів*** *(місяць-квартал-рік),* ***особливі відрізки часу*** *(п'яти- шести чи семиденний робочий тиждень),* ***сезонність*** *тощо.*

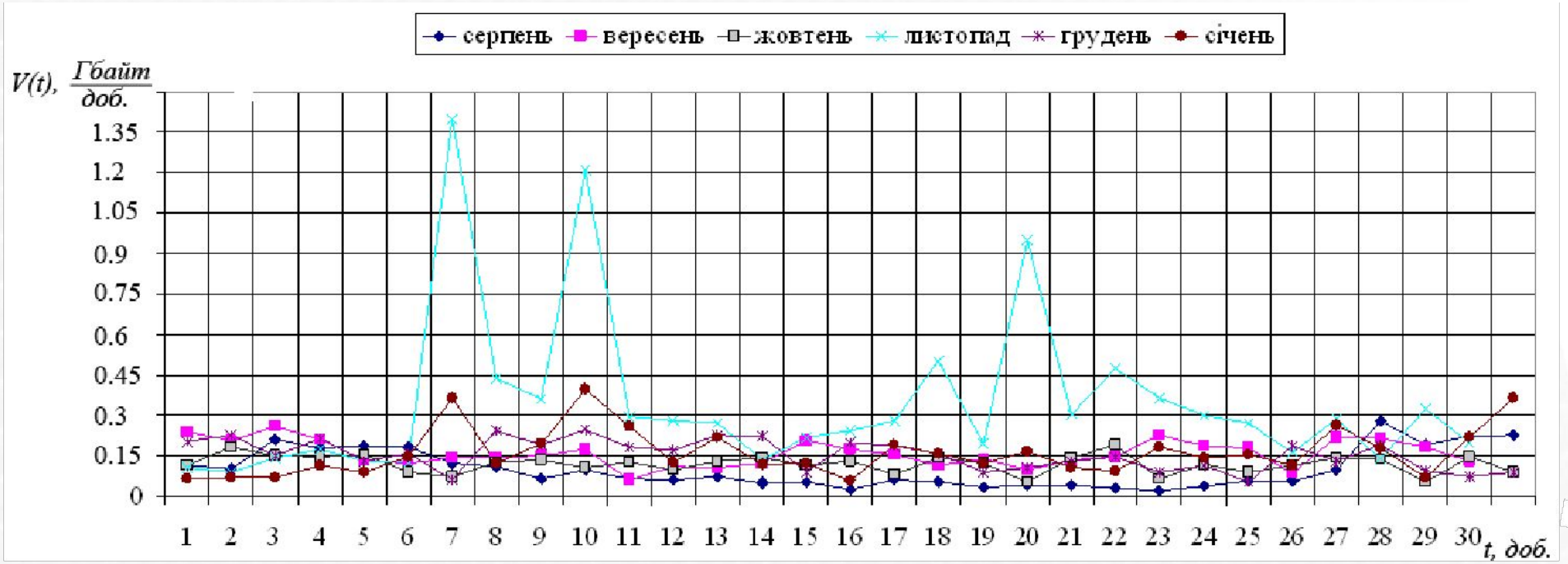
ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(передбачення поведінки системи в майбутньому)

ДОБА-ПІВРОКУ

Флуктуація середньодобового трафіка:
період спостереження з 01.08.2009 по 31.01.2010 року

ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(передбачення поведінки системи в майбутньому)

МІСЯЦЬ-ПІВРОКУ



Флуктуація місячного та піврічного трафіка:
період спостереження з 01.08.2009 по 31.01.2010 року

ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(постановка задачі)

ПОСТАНОВКА ЗАДАЧІ.

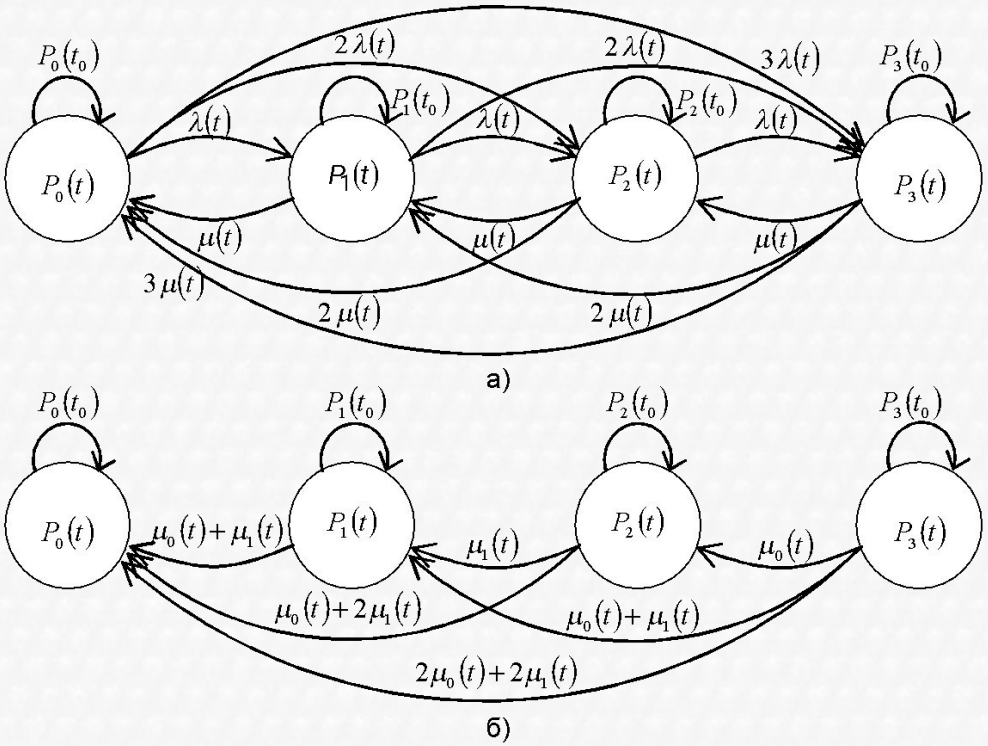
Вхідний сигнал

$$x(t) = s(t) + n(t), \qquad (1)$$

де $s(t)$ **- часова послідовність,** $t = \overline{0, N}$ **;** $n(t)$ **- шуми.**

Потрібно на основі «вікна» попереднього спостереження p **послідовності** $x(t)$ **знайти значення послідовності** $s(t)$ **в майбутньому, тобто** $s(t+l)$**-?, де** l **- крок прогнозування,** $l \in [1, N]$.

ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(постановка задачі)



Графова модель шаблонів поведінки Web-сервера Apache 2.2.10 (Linux|SUSE):
а – шаблон нормальної поведінки;
б – шаблон атаки

Модель шаблона нормальної поведінки:

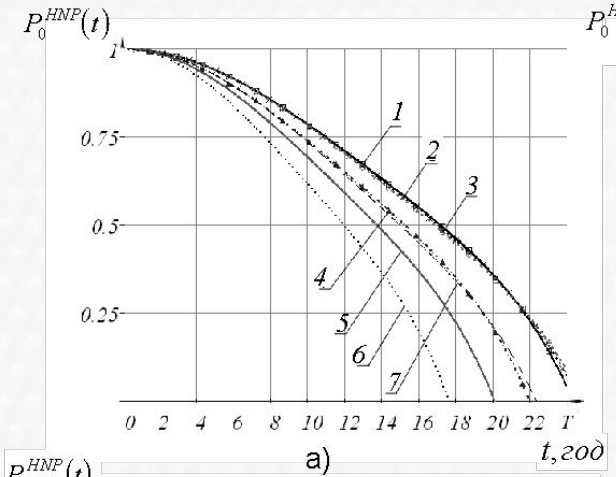
$$P_0^{HNP\,opt}(t) = 1 - 1.4469\left(t/T\right)^2 + 1.3649\left(t/T\right)^2 - 0.8747\left(t/T\right)^6.$$

Модель шаблона атаки:

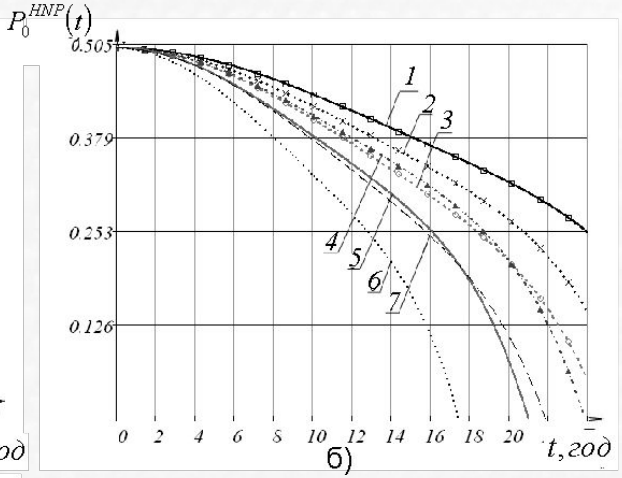
$$P_3^{HA}(t) = \frac{\left(16\mu_0^2 - 3\mu_1\right)^3}{16\mu_0^2\left(16\mu_0^2 - 9\mu_1\right)^2}\exp\left(\frac{-4\mu_0\left(16\mu_0^2 - 9\mu_1\right)}{16\mu_0^2 - 3\mu_1}t\right) +$$

$$+\left(1 - \frac{\left(16\mu_0^2 - 3\mu_1\right)^3}{16\mu_0^2\left(16\mu_0^2 - 9\mu_1\right)^2}\right)\exp\left(\frac{-4\mu_0 + \frac{\left(16\mu_0^2 - 3\mu_1\right)^2}{4\mu_0\left(16\mu_0^2 - 9\mu_1\right)}}{1 - \frac{\left(16\mu_0^2 - 3\mu_1\right)^3}{16\mu_0^2\left(16\mu_0^2 - 9\mu_1\right)^2}}t\right).$$

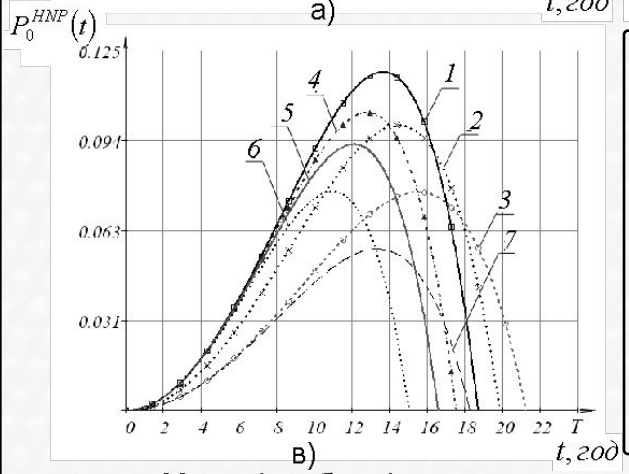
ПРОГНОЗУВАННЯ ЧАСОВИХ ПОСЛІДОВНОСТЕЙ
(постановка задачі)



а)



б)



в)

Моделі шаблонів нормальної поведінки *Web-сервера Apache 2.2.10 (Linux|SUSE)*: а – $P_0(t_0)=1,\ P_1(t_0)=P_2(t_0)=P_3(t_0)=0$;
б – $P_0(t_0)=0.5,\ P_1(t_0)=P_2(t_0)=0,\ P_3(t_0)=0.5$; в – $P_0(t_0)=P_1(t_0)=P_2(t_0)=0,\ P_3(t_0)=1$

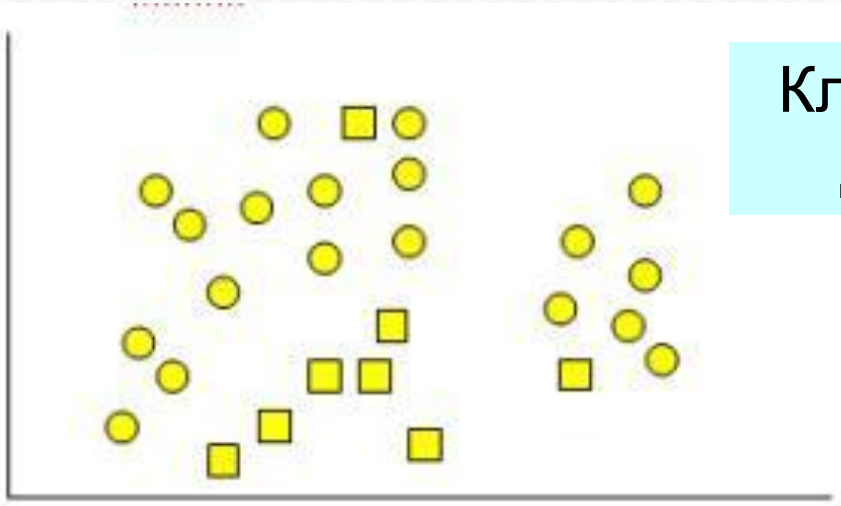
1 – $\lambda_{\min}^{opt}(t),\ \mu_{\max}^{opt}(t)$;
2 – $\lambda_{\min}^{opt}(t),\ \mu(t)=0.75\,\mu_{\max}^{opt}(t)$;
3 – $\lambda_{\min}^{opt}(t),\ \mu(t)=0.5\,\mu_{\max}^{opt}(t)$;
4 – $\lambda(t)=1.25\lambda_{\min}^{opt}(t),\ \mu_{\max}^{opt}(t)$;
5 – $\lambda(t)=1.5\lambda_{\min}^{opt}(t),\ \mu_{\max}^{opt}(t)$;
6 – $\lambda(t)=2\lambda_{\min}^{opt}(t),\ \mu_{\max}^{opt}(t)$;
7 – $\lambda(t)=1.25\lambda_{\min}^{opt}(t),\ \mu(t)=0.5\,\mu_{\max}^{opt}(t)$

Таблиця

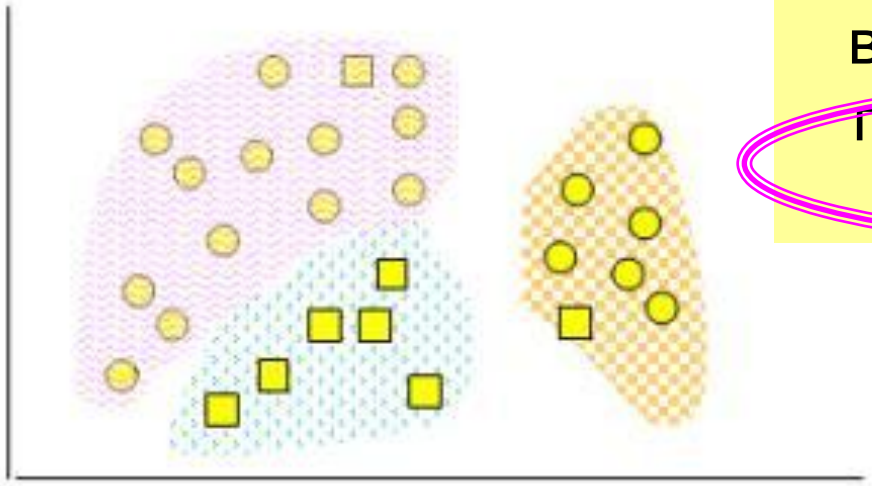
Вхідні дані для моделювання

Тривалість атаки, T год	б	Стратегії, $ãîä^{-1}$			Модель ША	Плата	Реакція СВА
		$\mu_0(t)$	$\mu_1(t)$	$\mu_2(t)$	$P_3^{HA}(t)$	I'	
		0.038	$0.074t$	$0.038+0.074t$	$-0.9e^{-0.48t}+1.9e^{-0.3t}$	0.11	Сигналізація про атаку
		0.048	$0.093t$	$0.048+0.093t$	$-0.6e^{-0.64t}+1.6e^{-0.36t}$	0.17	
		0.056	$0.11t$	$0.056+0.11t$	$-0.5e^{-0.76t}+1.5e^{-0.4t}$	0.32	Виявлення атаки сканування портів
		0.066	$0.13t$	$0.066+0.13t$	$-0.4e^{-0.92t}+1.4e^{-0.44t}$	0.61	Виявлення DoS-атаки
		0.076	$0.15t$	$0.076+0.15t$	$-0.3e^{-1.1t}+1.3e^{-0.48t}$	1	Виявлення DDoS-атаки

КЛАСТЕРИЗАЦІЯ
(розбиття даних на групи)



Класифікація – вихідні дані для класів – визначені.



Кластеризація – класи не визначені, здійснюється пошук найбільш схожих груп.

КЛАСТЕРИЗАЦІЯ
(постановка задачі)

Дано:

X **- множина об'єктів;**

Y **- множина кластерів.**

Відомо функцію відстані між об'єктами $\rho(x, x')$.

Задача – розбити X **на підмножини, що не перетинаються та віднести об'єкти** x_i **до відповідних кластерів** y_i .

КЛАСТЕРИЗАЦІЯ
(алгоритм розв'язку)

АЛГОРИТМ КЛАСТЕРИЗАЦІЇ

– **функція** $\alpha : X \to Y$, **що кожному** $x \in X$

ставить у відповідність номер кластера $y \in Y$.

КЛАСТЕРИЗАЦІЯ – навчання без вчителя.

АСОЦІАЦІЯ
*(знаходження **трендів** (однакових) ділянок)*

***ТРЕНД** (з англ. Trend - тенденція) - основна тенденція зміни часового ряду.*

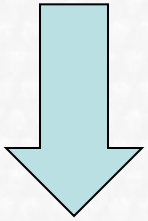
Тренди описують лінійними, логарифмічними, показниковими та ін. рівняннями.

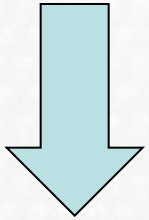
***ПРИНЦИП ВСТАНОВЛЕННЯ ТИПУ ТРЕНДА** – підбір його функціональної моделі статистичними методами або згладжуванням вихідного часового ряду.*

АСОЦІАЦІЯ
*(методи знаходження **трендів**)*

***Параметричні** – тренд описується гладкою функцією.*

***Непараметричні** – методи згладжування вихідного ряду.*





ЛІНЕАРИЗАЦІЯ НА ОСНОВІ МНК

ЕКСПОНЕНЦІАЛЬНЕ ЗГЛАДЖУВАННЯ

ПОСЛІДОВНІСТЬ

(знаходження нового знання в базах даних)

Числові **Функціональні**

АЛГОРИТМ ПОСЛІДОВНОСТІ

Послідовність – функція визначена на множині натуральних чисел, яка набуває значення на об'єктах довільної природи.

$$f : N \to X \ .$$

Записується у вигляді $\{x_n\}$.

Елементи x_n **називаються членами послідовності.**

3. Методи інтелектуальних обчислень.

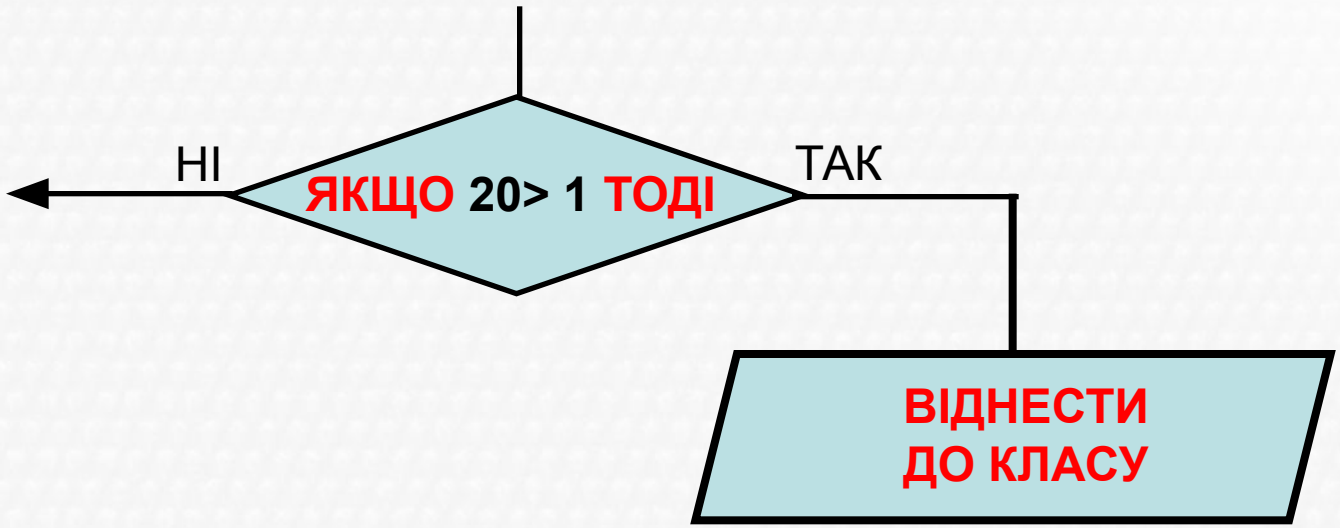
БАЗОВІ МЕТОДИ

- нейронні мережі;
- дерева рішень;
- системи міркування на основі аналогічних випадків;
- алгоритми визначення асоціацій і послідовностей;
- нечітка логіка;
- генетичні алгоритми;
- еволюційне програмування;
- візуалізація даних.

ДЕРЕВА РІШЕНЬ
(задачі чисельного прогнозу)

СУТЬ: створення ієрархічної структури **ЯКЩО** … **ТОДІ** у вигляді “дерева”.

ПРИКЛАД:



ПРОБЛЕМА ЗНАЧИМОСТІ

СИСТЕМИ МІРКУВАННЯ НА ОСНОВІ АНАЛОГІЧНИХ ВИПАДКІВ (“метод найближчого сусіда”)

СУТЬ: обчислення здійснюється за аналогом до ситуацій.

Перевага: застосовуються для обчислення при створенні різнотипних ІСЗІ.

Недолік: не створюють ні моделей ні правил; незрозумілість у виборі аналогів.

АЛГОРИТМИ ВИЗНАЧЕННЯ АСОЦІАЦІЙ І ПОСЛІДОВНОСТЕЙ

Алгоритми виявлення асоціацій знаходять правила для обчислення характеристик зображень, що поступають на вхід ІСЗІ.

Характеристики асоціацій

Довірчість

Потужність

Послідовність - це теж асоціація, але залежна від часу.

Це дозволяє працювати з серією подій для знаходження послідовних асоціацій протягом деякого періоду часу.

ЕВОЛЮЦІЙНЕ ПРОГРАМУВАННЯ
(формування гіпотез на мові програмування)

Суть – система знаходить програму, вносить в неї модифікації й здійснює обчислення.

Спеціальний транслюючий модуль, переводить знайдені залежності з внутрішньої мови системи на зрозумілу користувачу мову (математичні формули, таблиці тощо).

Метод групового урахування аргументів (МГУА) – залежності шукають у формі поліномів.

ВІЗУАЛІЗАЦІЯ ДАНИХ
(комбінування відомих методів обчислень)

СУТЬ: підвищення точності обчислень та їх швидкодії.

Програми візуалізації даних у певному сенсі не є засобом аналізу інформації, оскільки **вони тільки представляють її користувачу**.

Візуальне представлення виразно **узагальнює надвеликі обсяги даних.**