**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ**

**КАФЕДРА** **КОМП’ЮТЕРИЗОВАНИХ СИСТЕМ ЗАХИСТУ ІНФОРМАЦІЇ**

ДОПУСТИТИ ДО ЗАХИСТУ

Завідувач кафедри

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.В. Ільєнко

«\_\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ р.

На правах рукопису

УДК 004.056.5:510.22(043.3)

**МАГІСТЕРСЬКА АТЕСТАЦІЙНА РОБОТА**

**ВИПУСКНИКА ОСВІТНЬОГО СТУПЕНЯ**

**«МАГІСТР»**

**Тема**:Методи аналізу кіберзагроз в інформаційному просторі

|  |  |
| --- | --- |
| **Автор:** | Ю.І. Фіненко |
| **Науковий керівник:** к.т.н., доц. | Н.К. Гулак |
| **Нормоконтролер:** асист. | С.В. Єгоров |

**Київ 2020**

**НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ**

**Факультет:** Кібербезпеки, комп’ютерної та програмної інженерії

**Кафедра:** Компютеризованих систем захисту інформації

**Освітній ступінь:** Магістр

**Спеціальність:** 125 «Кібербезпека»

**Освітньо-професійна програма**: «Безпека інформаційних і комунікаційних систем»

ЗАТВЕРДЖУЮ

Завідувач кафедри

\_\_\_\_\_\_\_\_\_\_\_\_\_ С.В. Казмірчук

«\_\_» \_\_\_\_\_\_\_\_\_ 20\_\_ р.

**ЗАВДАННЯ**

**на виконання магістерської атестаційної роботи**

**магістранта Фіненка Юрія Івановича**

1. Тема: *Методи аналізу кіберзагроз в інформаційному просторі*

затверджена наказом ректора від 02.10.2019 р. № 2265/ст*.*

1. Термін виконання з 14.10.2019 р. по 09.02.2020 р*.*
2. Вихідні дані: вибірка контейнерів, які завідомо до початку дослідження були класифіковані на два різні класи: «порожній контейнер», «наповнений контейнер»; методи стеганографії та стеганоаналізу; методи математичної статистики; методи машинного навчання; програмне середовище «MATLAB».
3. Зміст пояснювальної записки: аналіз існуючих методів стеганоаналізу на основі нормативно-правового регулювання законодавства України; аналіз математичних моделей аналізу даних та каналів передачі прихованої інформації; створення алгоритму для оцінки наявності прихованої інформації в каналі зв’язку.

**КАЛЕНДАРНИЙ ПЛАН**

**виконання магістерської роботи**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Етапи виконання магістерської роботи** | **Термін виконання етапів** | **Примітка** |
|  | Уточнення постановки задачі | 14.10.2019 | *Виконано* |
|  | Аналіз літературних джерел | 15.10.2019-22.102019 | *Виконано* |
|  | Обґрунтування вибору рішення | 23.10.2019-30.10.2019 | *Виконано* |
|  | Збір інформації | 31.10.2019-15.11.2019 | *Виконано* |
|  | Дослідження існуючих сучасних методів аналізу кіберзагроз на основі правового-регулювання законодавства України | 15.11.2019-10.12.2019 | *Виконано* |
|  | Аналіз математичних моделей аналізу даних та каналів передачі прихованих кіберзагроз | 10.12.2019-01.01.2020 | *Виконано* |
|  | Розробка алгоритму для оцінки наявності прихованої кіберзагрози в каналі зв’язку | 02.01.2020-12.01.2020 | *Виконано* |
|  | Апробація роботи на науково-технічній конференції Університету «Україна» | 13.01.2020 | *Виконано* |
|  | Перевірка на антиплагіат | 02.02.2020 | *Виконано* |
|  | Оформлення і друк пояснювальної записки | 04.02.2020 | *Виконано* |
|  | Оформлення презентації | 05.02.2020 | *Виконано* |
|  | Отримання рецензій від рецензента | 08.02.2020 | *Виконано* |
|  | Захист в ЕК | 09.02.2020 | *Виконано* |

Магістрант Ю. Фіненко

(підпис, дата)

Науковий керівник Н. Гулак

(підпис, дата)

# РЕФЕРАТ

Магістерська атестаційна робота складається зі вступу, трьох розділів, загальних висновків, списку використаних джерел, додатків і має 122 сторінки основного тексту, 42 рисунка, 25 таблиць, 13 сторінок додатків. Список використаних джерел містить 20 найменування і займає 6 сторінок. Загальний обсяг роботи 135 сторінок.

Мета роботи полягає в обранні найпоширенішого методу аналізу кіберзагроз та в підвищенні ефективністі методу за рахунок розробки нового алгоритму.

В роботі проаналізовано існуючі методи аналізу кіберзагроз в інформаційному просторі було обрано найпоширеніший метод - статистичний, проаналізовано математичні моделі аналізу даних та канали управління кібератаками, розроблено алгоритм, який покращує обраний метод аналізу даних для підвищення ефективності використання обраного методу.

Розроблений алгоритм дає можливість пристосування технологій машинного навчання для захисту інформаційного простору завдяки аналізу первинних даних на предмет прихованих кіберзагроз.

Можливі напрямки розвитку цієї роботи пов’язані із розширенням моделі і алгоритму та розробка прикладного програмно-технічного забезпечення відповідно до вимог міжнародних стандартів, для більш повноцінного аналізу прихованих кіберзагроз в існуючих та перспективних каналів зв’язку.

Ключові слова: МАШИННЕ НАВЧАННЯ, КІБЕРЗАГРОЗА, ІНФОРМАЦІЙНИЙ ПРОСТІР, ЛІНІЯ ЗВ’ЯЗКУ, МАТЕМАТИЧНА СТАТИСТИКА, АТАКА НА КІБЕРЗАГРОЗУ, МАТЕМАТИЧНА МОДЕЛЬ, БІНАРНА КЛАСИФІКАЦІЯ КАНАЛІВ ЗВ’ЯЗКУ.

**ЗМІСТ**

[РЕФЕРАТ 4](#_Toc30500867)

[ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ 7](#_Toc30500868)

[ВСТУП 8](#_Toc30500869)

[Розділ 1. Кіберзагрози та методи їх аналізу 12](#_Toc30500870)

[1.1. Визначення проблеми. 12](#_Toc30500871)

[1.3. Основні поняття та визначення 13](#_Toc30500872)

[1.3. Модель загроз інформаційному просторі 17](#_Toc30500873)

[1.4. Модель порушника 25](#_Toc30500874)

[1.5. Сучасні кіберзагрози інформаційному простору 29](#_Toc30500875)

[1.6. Аналіз кіберзагроз 34](#_Toc30500876)

[1.7. Система управління інформаційною безпекою 37](#_Toc30500877)

[1.8. Висновок до розділу 40](#_Toc30500878)

[Розділ 3. Аналіз прихованої інформації, машинне навчання 41](#_Toc30500879)

[3.1. Класифікація методів стеганографії 41](#_Toc30500880)

[3.3. Класифікація методів стегоаналізу 42](#_Toc30500881)

[3.3.1. Стеганоаналіз методом критерію правдоподібності Хі-квадрат 45](#_Toc30500882)

[3.3. Машинне навчання 47](#_Toc30500883)

[3.3.1. Поняття машинного навчання 47](#_Toc30500884)

[3.3.3. Види машинного навчання 47](#_Toc30500885)

[3.3.3. Алгоритми машинного навчання 48](#_Toc30500886)

[3.3.4. Види регресійних моделей 48](#_Toc30500887)

[3.3.5. Метод найменших квадратів 51](#_Toc30500888)

[3.3.6. Інтерполяція і екстраполяція. 52](#_Toc30500889)

[3.3.7. Обчислення сум квадратів. 52](#_Toc30500890)

[3.3.9. Середньоквадратична помилка оцінки 54](#_Toc30500891)

[3.3.10. Припущення 55](#_Toc30500892)

[3.3.11. Аналіз залишків 56](#_Toc30500893)

[3.3.13. Оцінка придатності емпіричної моделі 56](#_Toc30500894)

[3.3.14. Перевірка гіпотез про нахил і коефіцієнті кореляції 58](#_Toc30500895)

[3.3.14. Оцінка математичного очікування і передбачення індивідуальних значень 61](#_Toc30500896)

[3.3.15. Критерій згоди «-квадрат» 67](#_Toc30500897)

[3.3.17. Класифікація на основі Байєсівського підходу 68](#_Toc30500898)

[3.3.17 Наївний байесовский класифікатор 69](#_Toc30500899)

[3.3.19. Поняття кластеризації 70](#_Toc30500900)

[3.3.30. Класифікація алгоритмів кластеризації 72](#_Toc30500901)

[3.3.31. Нейронні мережі Кохонена 74](#_Toc30500902)

[3.3.33. Кластеризація методом k-means 80](#_Toc30500903)

[3.3.34. EМ - масштабований алгоритм кластеризації 83](#_Toc30500904)

[3.4. Висновок до розділу 89](#_Toc30500905)

[Розділ 3. Алгоритм оцінки наявності прихованої інформації в каналі зв’язку 91](#_Toc30500906)

[3.1. Причини використання стеганографії 91](#_Toc30500907)

[3.3. Статистична оцінка двох класів контейнерів 94](#_Toc30500908)

[3.3. Принци роботи системи стеганоаналізу на основі розпізнавання образів та машинного навчання 103](#_Toc30500909)

[ВИСНОВКИ 104](#_Toc30500910)

[СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ 105](#_Toc30500911)

[ДОДАТКИ 112](#_Toc30500912)

# ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

# ВСТУП

**Актуальність.** Так, протягом останніх років на інформаційно-телекомунікаційні системи деяких об’єктів, які за своїм значенням і роллю для життєдіяльності суспільства є об’єктами критичної інфраструктури, здійснено низку масштабних кібератак, зокрема:

* 21 - 25 травня 2014 відбулися DDoS-атаки і злом сайту ЦВК під час президентських виборів, внаслідок яких на сайті з’явилися помилкові результати. Незважаючи на повідомлення про злом, саме ці дані були озвучені в новинах на російському Першому каналі як реальні результати виборів в Україні [1];
* у червні 2014 року на серверах приватних компаній України і країн НАТО були виявлені шкідливі програми, які займалися кібершпіонажем. Серед них такі, як Turla/Uroburos/Snake, RedOctober, MiniDuke і NetTraveler [1];
* 23 грудня 2015 року за допомогою троянської програми BlackEnergy3, у використанні якої були раніше помічені російські хакери, було відключено близько 30 підстанцій Прикарпаттяобленерго, в зв'язку з чим більше ніж 200 тисяч жителів Івано-Франківської області залишалися без електроенергії на термін від одного до п’яти годин. Тоді ж відбулися атаки на Київобленерго і Чернівціобленерго [1];
* 6 грудня 2015 року відбулася хакерська атака на внутрішні телекомунікаційні мережі Мінфіну, Держказначейства, Пенсійного фонду, що вивела з ладу ряд комп'ютерів, а також знищила критично важливі бази даних, що призвело до затримки бюджетних виплат на сотні мільйонів гривень [1];
* 15 грудня 2015 року українські хакери на замовлення невстановленої особи із Санкт-Петербурга здійснили DDOS-атаку на сайт Укрзалізниці, внаслідок чого протягом дня була повністю заблокована його робота. Атака була націлена на крадіжку даних про пассажироперевезення [1];
* 17 грудня 2015 року кібератака на підстанцію “Північна” компанії “Укренерго” призвела до збою в автоматиці управління, через що більше години знеструмленими залишалися райони у північній частині правобережного Києва і прилеглі райони області [1];
* у першій половині дня 27 червня 2017 року розпочалася масова кібератака на український державний та комерційний сектор із застосування шкідливого програмного забезпечення – вірусу-шифрувальника файлів Petya Ransomware. Її жертвами стали інформаційно-телекомунікаційні системи “Укрпошти”, аеропорту “Бориспіль”, “Укренерго”, ДТЕК, багатьох банків, ЗМІ, телеканалів, АЗС та інших компаній [1].

У комп'ютерній безпеці різні методи машинного навчання давно застосовуються в фільтрації спаму, аналізі трафіку, при виявленні фроду або шкідливого програмного забезпечення. Однак є ще одна не менш важлива задача в сфері захисту інформації, це викриття прихованих кіберзагроз, які можуть завдати непоправної шкоди інформаційно-комунікаційним системам, які здійснюються через відкриті канали зв'язку ззовні, за допомогою методів приховування факту передачі. Аналіз і виявлення таких кіберзагроз та каналів управління ними є пріоритетними напрямками сфери кібербезпеки. Виділити з мільярдів сигналів і величезних масивів різноформатних даних інформацію, яка реально важлива для відбиття атаки, вкрай складно. Людина витратить на такий аналіз занадто багато часу. І, навпаки, система машинного навчання може проводити поведінковий аналіз мільярдів кіберінцидентів кожен день. Це дозволяє значно скоротити час реагування на кіберінциденти. Тому дане дослідження є пріоритетним і затребуваним сьогодні [].

Використання методів аналізу кіберзагроз та машинного навчання дозволить ефективніше виявляти приховані кіберзагрози інформаційним системам. Тому розробка нових аналітичних методів для захисту інформації є актуальною науково-практичною задачею [3].

Для аналізу дослідження були розглянуті методи математичної статистики, за якими було показано різницю між різними каналами передачі кіберзагроз в інформаційному просторі [3].

**Мета роботи.** На підставі ретельного аналізу методів кіберзагроз обрати найпоширеніший метод аналізу та підвищити ефективність методу за рахунок розробки нового алгоритму.

Для забезпечення поставленої мети, потрібно виконати ряд завдань:

* проаналізувати існуючі методи аналізу кіберзагроз в інформаційному просторі на основі нормативно-правового регулювання законодавства України;
* за категоріями жертв проаналізувати математичні моделі аналізу даних та канали управління кібератаками;
* розробка алгоритму для підвищення ефективності використання обраного методу.

**Галузь застосування.** Розроблений алгоритм можливо використовувати в галузі кібербезпеки для підвищення ефективності існуючих методів та систем аналізу кіберзагроз.

**Об’єктом дослідження** є процес аналізу та оцінки прихованих кіберзагроз в інформаційному просторі.

**Предметом дослідження** є канал зв’язку з прихованими кіберзагрозами.

**Методи дослідження.** Для вирішення означених вище наукових завдань в роботі використані методи системного аналізу, теорії інформаційної безпеки, методи аналізу кіберзагроз, методи математичної статистики та машинного навчання.

**Наукова новизна отриманих результатів:** На основі поєднання сучасних методів захисту інформації, математичної статистики та математичних моделей аналізу даних було розроблено алгоритм, що дозволило удосконалити технологію оцінки та бінарної класифікації відкритих каналів зв’язку з прихованими кіберзагрозами.

**Практичне значення отриманих результатів:** Результатом дослідження дає можливість пристосування технологій машинного навчання для захисту інформаційно-комунікаційних систем завдяки аналізу первинних даних на предмет кіберзагроз.

**Апробація результатів.**

1. Шматок А.С. Методы анализа критических данных на основе машинного обучения. / А.С. Шматок, Ю.И. Финенко // ОРАЛДЫҢ ҒЫЛЫМ ЖАРШЫСЫ - №3 (130) 3019. - Оралқаласы, ЖШС «Уралнаучкнига», 3019. – С. 58-63.
2. Шматок А.С. Методы анализа критических данных на основе машинного обучения. / А.С. Шматок, Ю.И. Финенко // ІІ міжнар. наук.-практ. конф. «Проблеми кібербезпеки інформаціно-телекомунікаційних систем (PCSITS)», «Київ, 11-13 квітня 3019 р.). – Київ, КНУ ім. Тараса Шевченка, 3019. – С. 31-33.
3. Шматок О.С. Штучний інтелект та машинне навчання в задачах стеганоаналізу даних. / О.С. Шматок, Ю.І. Фіненко, А.Б. Єлізаров, В.А. Телющенко // Вісник Університету «Україна», Серія: «Інформатика, обчислювальна техніка та кібернетика» - №3 (33) 3019. – Київ, Університет «Україна», 3019. С.319-337.

# Розділ 1. Кіберзагрози та методи їх аналізу

## 1.1. Визначення проблеми.

Відповідно до частини третьої статті 6 Закону України “Про основні засади забезпечення кібербезпеки України” Адміністрацією Держспецзв’язку розроблено проект постанови Кабінету Міністрів України “Деякі питання проведення незалежного аудиту інформаційної безпеки на об’єктах критичної інфраструктури” (далі – проект постанови) [4].

Стратегією кібербезпеки України, затвердженою Указом Президента України від 15.03.3015 № 96, визначено основні загрози кібербезпеці, зокрема для об'єктів критичної інфраструктури, шляхи протидії їм та зазначено, що сучасні інформаційно-комунікаційні технології можуть використовуватися для вчинення терористичних актів [4].

Аналіз кіберзагроз свідчить, що кібератаки на комунікаційні системи та системи управління технологічними процесами об’єктів критичної інфраструктури держави таких галузей, як енергетика, хімічна промисловість та інші можуть призвести до виникнення надзвичайних ситуацій техногенного характеру та/або негативного впливу на стан екологічної безпеки держави [4].

З урахуванням потреб національної безпеки і необхідності запровадження системного підходу до розв’язання проблеми на загальнодержавному рівні створення системи захисту критичної інфраструктури є одним із пріоритетів у реформуванні сектору оборони і безпеки України [4].

Водночас Закон України “Про основні засади забезпечення кібербезпеки України” визначає, що до Переліку об’єктів критичної інфраструктури (далі – Перелік) можуть бути віднесені підприємства, установи та організації незалежно від форми власності, які провадять діяльність та надають послуги в галузях енергетики, хімічної промисловості, транспорту, інформаційно-комунікаційних технологій, електронних комунікацій, у банківському та фінансовому секторах; надають послуги у сферах життєзабезпечення населення, зокрема у сферах централізованого водопостачання, водовідведення, постачання електричної енергії і газу, виробництва продуктів харчування, сільського господарства, охорони здоров’я; є комунальними, аварійними та рятувальними службами, службами екстреної допомоги населенню; включені до переліку підприємств, що мають стратегічне значення для економіки і безпеки держави; є об’єктами потенційно небезпечних технологій і виробництв [4].

Необхідність прийняття постанови зумовлена відсутністю відомостей щодо реального стану інформаційної безпеки на об’єктах критичної інфраструктури та, як наслідок, унеможливлює системний підхід до розв’язання проблеми захисту критичної інфраструктури на загальнодержавному рівні [4].

Проблеми забезпечення належного рівня інформаційної безпеки на об’єктах критичної інфраструктури не можуть бути розв’язані без існування систематизованого підходу до аналізу стану захисту інформації, який базувався би на реальних показниках, отриманих під час проведення незалежного аудиту інформаційної безпеки [4].

Основні групи (підгрупи), на які проблема впливає:

* Громадяни
* Держава
* Суб’єкти господарювання у тому числі суб’єкти малого підприємництва [4].

## 1.3. Основні поняття та визначення

Визначені правові і організаційні засади забезпечення захисту національних інтересів України в кіберпросторі, основні цілі, напрями і принципи державної політики у сфері кібербезпеки, повноваження державних органів, підприємств, установ, організацій, осіб і громадян в цій сфері, основні принципи координації їх діяльності по забезпеченню кібербезпеки [4].

Індикатори кіберзагроз - показники (технічні дані), що використовуються для виявлення та реагування на кіберзагрози [4].

Інформація про інцидент кібербезпеки - відомості про обставини кіберінциденту, зокрема про те, які об’єкти кіберзахисту і за яких умов зазнали кібератаки, які з них успішно виявлені, нейтралізовані, яким запобігли за допомогою яких засобів кіберзахисту, у тому числі з використанням яких індикаторів кіберзагроз [4].

Інцидент кібербезпеки (далі - кіберінцидент) - подія або ряд несприятливих подій ненавмисного характеру (природного, технічного, технологічного, помилкового, у тому числі внаслідок дії людського фактора) та/або таких, що мають ознаки можливої (потенційної) кібератаки, які становлять загрозу безпеці систем електронних комунікацій, систем управління технологічними процесами, створюють імовірність порушення штатного режиму функціонування таких систем (у тому числі зриву та/або блокування роботи системи, та/або несанкціонованого управління її ресурсами), ставлять під загрозу безпеку (захищеність) електронних інформаційних ресурсів [4].

Кібератака - спрямовані (навмисні) дії в кіберпросторі, які здійснюються за допомогою засобів електронних комунікацій (включаючи інформаційно-комунікаційні технології, програмні, програмно-апаратні засоби, інші технічні та технологічні засоби і обладнання) та спрямовані на досягнення однієї або сукупності таких цілей: порушення конфіденційності, цілісності, доступності електронних інформаційних ресурсів, що обробляються (передаються, зберігаються) в комунікаційних та/або технологічних системах, отримання несанкціонованого доступу до таких ресурсів; порушення безпеки, сталого, надійного та штатного режиму функціонування комунікаційних та/або технологічних систем; використання комунікаційної системи, її ресурсів та засобів електронних комунікацій для здійснення кібератак на інші об’єкти кіберзахисту [4].

Кібербезпека - захищеність життєво важливих інтересів людини і громадянина, суспільства та держави під час використання кіберпростору, за якої забезпечуються сталий розвиток інформаційного суспільства та цифрового комунікативного середовища, своєчасне виявлення, запобігання і нейтралізація реальних і потенційних загроз національній безпеці України у кіберпросторі [4].

Кіберзагроза - наявні та потенційно можливі явища і чинники, що створюють небезпеку життєво важливим національним інтересам України у кіберпросторі, справляють негативний вплив на стан кібербезпеки держави, кібербезпеку та кіберзахист її об’єктів. Кіберзахист - сукупність організаційних, правових, інженерно-технічних заходів, а також заходів криптографічного та технічного захисту інформації, спрямованих на запобігання кіберінцидентам, виявлення та захист від кібератак, ліквідацію їх наслідків, відновлення сталості і надійності функціонування комунікаційних, технологічних систем [4].

Кіберзлочин (комп’ютерний злочин) - суспільно небезпечне винне діяння у кіберпросторі та/або з його використанням, відповідальність за яке передбачена законом України про кримінальну відповідальність та/або яке визнано злочином міжнародними договорами України [4].

Кіберзлочинність - сукупність кіберзлочинів [4].

Кібероборона - сукупність політичних, економічних, соціальних, військових, наукових, науково-технічних, інформаційних, правових, організаційних та інших заходів, які здійснюються в кіберпросторі та спрямовані на забезпечення захисту суверенітету та обороноздатності держави, запобігання виникненню збройного конфлікту та відсіч збройній агресії [4].

Кіберпростір - середовище (віртуальний простір), яке надає можливості для здійснення комунікацій та/або реалізації суспільних відносин, утворене в результаті функціонування сумісних (з’єднаних) комунікаційних систем та забезпечення електронних комунікацій з використанням мережі Інтернет та/або інших глобальних мереж передачі даних [4].

Кіберрозвідка - діяльність, що здійснюється розвідувальними органами у кіберпросторі або з його використанням [4].

Кібертероризм - терористична діяльність, що здійснюється у кіберпросторі або з його використанням [4].

Кібершпигунство - шпигунство, що здійснюється у кіберпросторі або з його використанням. Критична інформаційна інфраструктура - сукупність об’єктів критичної інформаційної інфраструктури [4].

Критично важливі об’єкти інфраструктури (далі - об’єкти критичної інфраструктури) - підприємства, установи та організації незалежно від форми власності, діяльність яких безпосередньо пов’язана з технологічними процесами та/або наданням послуг, що мають велике значення для економіки та промисловості, функціонування суспільства та безпеки населення, виведення з ладу або порушення функціонування яких може справити негативний вплив на стан національної безпеки і оборони України, навколишнього природного середовища, заподіяти майнову шкоду та/або становити загрозу для життя і здоров’я людей [4].

Кіберзахисту підлягають інформаційно-комунікаційні системи усіх форм власності, в яких обробляються національні інформаційні ресурси і які використовуються в інтересах органів державної влади і місцевого самоврядування, правоохоронних органів і військових формувань, в сферах електронного управління, електронних державних послуг, електронної комерції, електронного документообігу, а також об'єкти критичної інформаційної інфраструктури. Перелік вказаних об’єктів має затверджуватися Кабінетом Міністрів України та наразі відсутній. До останніх можуть бути віднесені підприємства, установи і організації незалежно від форми власності: в області енергетики, хімічної промисловості, транспорту, інформаційно-комунікаційних технологій, електронних комунікацій, у банківському і фінансовому секторах; у сферах водо-, газо- і електропостачання, водовідведення, виробництва продуктів харчування, сільського господарства, охорони здоров'я. Також до об'єктів критичної і відносяться комунальні, аварійні і рятувальні служби, стратегічні підприємства, потенційно небезпечні виробництва [4].

Слід зазначити, що дія закону № 3153 не поширюється, зокрема, на: відносини та послуги, пов’язані із змістом інформації, що обробляється (передається, зберігається) у комунікаційних та/або в технологічних системах; діяльність, пов’язану із захистом інформації, що становить державну таємницю, комунікаційні та технологічні системи, призначені для її оброблення; соціальні мережі, приватні електронні інформаційні ресурси в мережі Інтернет (включаючи блог-платформи, відеохостинги, інші веб-ресурси), якщо такі інформаційні ресурси не містять інформації, необхідність захисту якої встановлено законом, відносини та послуги, пов’язані з функціонуванням таких мереж і ресурсів; комунікаційні системи, які не взаємодіють із публічними мережами електронних комунікацій (електронними мережами загального користування), не підключені до мережі Інтернет та/або інших глобальних мереж передачі даних (крім технологічних систем) [4].

## 1.3. Модель загроз інформаційному просторі

Діяльність всіх осіб, що мають доступ до охоронюваним законом масивів особистої інформації, повинна спиратися на норми, введені «Моделлю». Модель бачить чотири види суб'єктів, які можуть завдати шкоди охоронюваним законом інтересам [5]:

* громадяни, які мають на власні цілі;
* іноземні спецслужби;
* інші іноземні організації, терористичні або комерційні;
* кримінальні угруповання [5].

Їх кроки, які посягають на збереження і цілісність захищених масивів відомостей, можуть принести шкоду інтересам особистості, суспільства, держави [5].

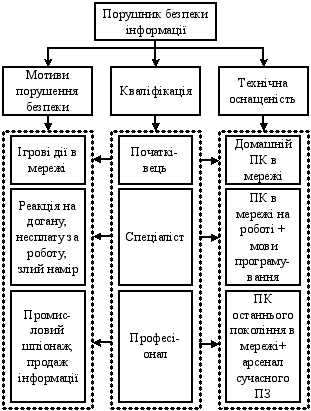


Рис. Суб’єкти, які можуть завдати шкоди інформаційним системам

ДССЗІ розробило модель дій цих типів суб'єктів. Зазіхання, за версією документа, можуть відбуватися наступними шляхами [5]:

* перехоплення або знімання інформації, яка направляється по каналах зв'язку, для її копіювання або поширення з будь-якими цілями, такими, що суперечать законодавству [5];
* отримання неправомірного доступу до баз, в яких зберігаються номери паспортів, адреси, медичні історії. Доступ використовується не тільки для копіювання або неправомірного поширення відомостей, а й для їх зміни, знищення, внесення спотворень в важливі параметри. Для деструктивних впливів використовуються спеціальні програмні та технічні засоби, при цьому оператори часто не мають готових відповідей на виклики, створені з використанням сучасних технологій [5].

ДССЗІ, розробляючи свою модель загроз, було впевнено в тому, що з її допомогою кожен конкретний оператор зможе розробити власні методики захисту від загроз в інформаційному просторі власних підприємств. Для цього потрібно правильно використовувати матеріальні засоби зберігання і передачі, серед яких відомство називає [5]:

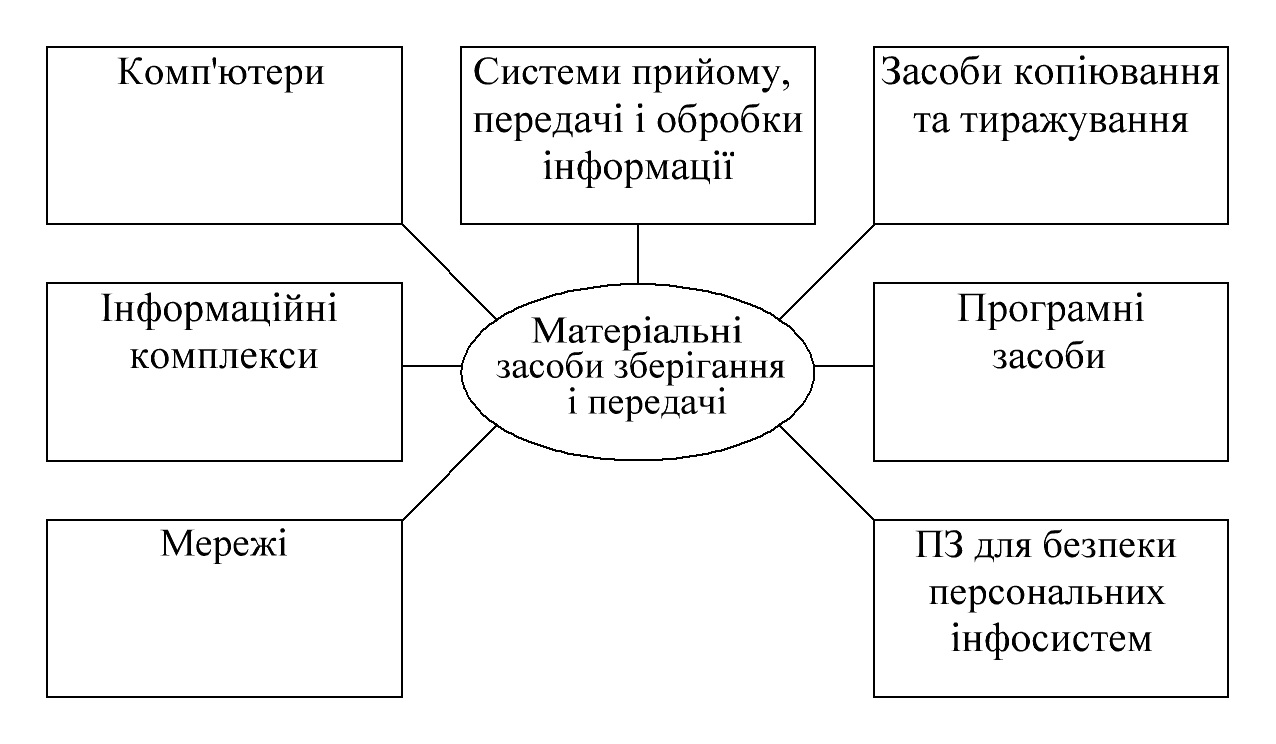


Рис. 1. Матеріальні засоби зберігання і обробки, яке необхідно захищати.

Відповідальність за працездатність і захищеність обладнання і мереж несуть їх власники [5].

Завдання, які вирішуються «Моделлю загроз»

Будь-який суб'єкт ринку, який потрапив до реєстру, повинен не тільки в будь-який момент бути готовим пройти перевірку ДССЗІ, але і вирішувати завдання, не передбачені чинною на поточний момент нормативною документацією, що стають новими викликами. До таких завдань відносяться [5]:

* розробка власних моделей загроз, актуальних для кожного конкретного особи, яка здійснює обробку персональних даних. Абсолютно різні цілі можуть бути в осіб, які зазіхають на інформацію, пов'язану з особистостями державних службовців, і тих злочинців, які цікавляться медичними картами пенсіонерів [5];
* аналіз і моніторинг поточної ситуації із захистом баз персональних даних від зовнішнього проникнення [5];
* розробка власної системи захисту баз даних і нейтралізації потенційних загроз, заснованої на пропонованих для кожної окремої системи методах і рекомендаціях [5];
* систематичне проведення заходів, які повинні виключити несанкціоноване проникнення до баз даних порушників - третіх осіб, яким не оформлений допуск до вивчення і обробці цієї інформації [5];
* організація технічної системи захисту засобів обробки даних, яка повинна виключити будь-який вплив на них, що може спричинити їх пошкодження або руйнування [5];
* постійний контроль якості захисту [5].

Модель загроз пропонує вивчити основні вразливі точки систем захисту, актуальні види загроз, доступні способи забезпечення інформаційної безпеки. UA-CERT проводить постійний моніторинг знову з'являються видів загроз, вносячи зміни в свої рекомендації [5].

Класифікація загроз

Неможливо боротися з будь-якими загрозами, якщо не розуміти їх походження, ступеня серйозності, інших параметрів. Вирішенню цього завдання допоможе запропонована класифікація загроз. Відомство в моделі пропонує наступне визначення загрози: це сукупність зовнішніх і внутрішніх факторів або умов, які дозволяють зловмисникам навмисно або ненавмисно отримати доступ до охоронюваної інформації, завдяки чому стає можливим її несанкціоноване [5]:

* знищення (стирання з пам'яті комп'ютера);
* блокування доступу до неї;
* поширення серед обмеженого або широкого кола осіб;
* зміна;
* копіювання;
* будь-які інші дії, не передбачені встановленими правилами обробки відомостей в інформаційних системах персональних даних [5].

Ризики можуть носити як особистісний, пов'язаний з наміром третіх осіб, так і знеособлений, техногенний, характер, пов'язаний з помилками або необережністю обслуговуючого персоналу. Сукупність факторів і умов для кожної окремої системи матиме індивідуальний характер, який складається з характеристик конкретної ІС, середовища та шляхів поширення інформаційних сигналів, технічних та інших можливостей, наявних у джерел потенційних загроз [5].

Характеристики ІС, які впливають на появу нових загроз і ризиків [5]:

* обсяг і зміст відомостей, наявних в базі;
* структура та конфігурація системи;
* наявність підключення системи до мереж передачі інформації загального користування або мереж, за якими проводиться транскордонне спілкування;
* наявність і якість систем захисту;
* режим, в якому обробляються персональні дані;
* наявність різнорівневої системи доступу осіб з різним обсягом повноважень і завдань до даних;
* фізичне місцезнаходження технічних пристроїв, режим їх захисту від протиправного посягання [5].

При цьому властивості середовища поширення сигналів враховуються з різних точок зору, часто отримати доступ до охоронюваним відомостями можна, використовуючи лінії підключення до електричних мереж, шляхом знімання інформації про зміну електромагнітного випромінювання [5].

Повністю досі не виключена можливість підслуховування і збереження звукової інформації за допомогою пристроїв звукозапису [5].

При аналізі джерел загроз необхідно вивчити такі їх типи, як можливості, що виникають у зв'язку з використанням службового становища, технічні, програмні, фізичні. Будь-яка загроза отримує свою реалізацію на практиці в той момент, коли між суб'єктом, що має намір здійснити протиправну дію, і об'єктом, що охороняється виникає канал доступу, що призводить до можливості виникнення несанкціонованого, усвідомленого або випадкового доступу до об'єкта [5].

Аналізуючи загрози, що виникають при обробці персональних даних, спираючись на багатогранність можливостей отримання доступу до інформації, модель виділяє такі їх класи [6]:

* внутрішні, інсайдерські – ризики викрадення або зміни відомостей виникають при реалізації можливостей внутрішнього користувача, доступ до інформаційних систем він отримує внаслідок свого службового становища, роботи в компанії або держструктурі [6];
* зовнішні - суб'єкт загроз отримує можливість несанкціонованого доступу до об'єкта захисту, використовуючи можливості мереж загального користування [6];
* технічні - виникають при використанні можливостей апаратних закладок і програм, призначених для розкрадання електронних відомостей [6].

Всі ці типи загроз вимагають серйозного аналізу особистостей співробітників, що мають доступ до чутливої інформації, і підвищеної уваги до технічної захищеності засобів зберігання і обробки даних з використанням спеціалізованих програмних засобів, DLP-систем, SIEM-систем, а також засобів захисту інформаційних ресурсів [6].

Також в моделі дається класифікація, заснована на різних типах використовуваного обладнання [6]:

* загрози безпеці тим даних, які обробляються співробітником на його робочому місці (АРМ), що не підключеному до мереж Інтернет [6];
* загрози відомостями, які обробляються співробітником на його робочому місці, підключеному до мереж Інтернет [6];
* погрози тим інформаційним масивам, які обробляються в локальних мережах, які не підключені до мереж Інтернет [6];
* погрози тим масивам, з якими працюють в локальних мережах підприємств і організацій, що мають вихід в Інтернет [6];
* загрози безпеці тих даних, які обробляються в розподілених мережах операторів і підключених, і не підключених до мереж міжнародного інформаційного обміну (Інтернет) [6].

Також в моделі загроз безпеки пропонується класифікація, пов'язана з різними типами технічних засобів, завдяки яким здійснюється доступ до охоронюваним масивів інформації [6]:

* використання шкідливих програм, вірусів, хробаків і аналогічних, що створюються з свідомо протиправними цілями;
* втрати даних з технічних і фізичних каналах витоку;
* інших спеціальних впливів [6].

За типами уразливості засобів захисту і технічних засобів, встановлених у конкретного оператора, модель загроз виділяє наступні [6]:

* пов'язані з уразливістю системного програмного забезпечення [6];
* пов'язані з використанням недоліків прикладного ПО [6];
* пов'язані з можливостями застосування апаратних закладок [6];
* пов'язані з використанням засобів зв'язку та протоколів передачі інформації [6];
* пов'язані із застосуванням технічних каналів передачі даних (наприклад, телефонних мереж, мереж електроживлення) [6].

Для виявлення цих точок уразливості необхідне проведення регулярного аудиту безпеки і систем захисту інформації. Крім аналізу можливостей доступу передбачуваних порушників режиму конфіденційності до засобів комп'ютерної обробки відомостей, необхідно мати на увазі, що практично будь-який офісне технічний засіб може стати потенційним джерелом загрози. Модель безпеки класифікує їх так [6]:

* АРМ;
* виділені засоби обробки (принтери, копіювальні апарати, звукозаписна апаратура, засоби відеоспостереження;
* мережі зв’язку [6].

Технічні канали витоку

Практика показує, що більша частина території, що охороняється інформації витікає з технічних каналам передачі даних. Сигнал поширюється в певній фізичному середовищі, він може бути акустичним або електромагнітним, його перехоплення здійснюється за допомогою закладних пристроїв і іншими шляхами. Пристрої можуть перехоплювати дані електромагнітного випромінювання, акустичну та візуальну інформацію. Захист від цього способу перехоплення здійснюється шляхом обмеження доступу на об'єкт, що охороняється [6].

Несанкціонований доступ

Дуже рідко організація може повністю захистити дані від дій власних співробітників. Для заподіяння шкоди даними можуть використовуватися [6]:

* штатне програмне забезпечення, що дозволяє потрапити в операційну середу;
* створення позаштатних умов роботи, що дозволяє використовувати одержувані спотворення для модифікації даних;
* шкідливі програми;
* загрози, пов'язані з віддаленим доступом;
* комбіновані загрози [6].

Для боротьби з ними необхідні установка захисного ПЗ, регулярний моніторинг функціонування робочих станцій. Крім того, ДССЗІ та UA-CERT проводять перевірки відповідності готовності технічних систем операторів і видає розпорядження, що дозволяють збільшити ступінь захищеності. Не завжди для цього потрібні серйозні фінансові кошти. Стандартні заходи безпеки і дотримання інструкцій допомагають забезпечити до 90% збереження даних [6].

## 1.4. Модель порушника

Якщо існує інформаційна система, у якій циркулює інформація з обмеженим доступом та конфіденційні дані, то знайдеться особа (порушник), метою якої буде ознайомлення з інформацією, її модифікація чи знищення. Для того, щоб розробити комплекс заходів по забезпеченню захищеності інформаційних ресурсів, необхідно побудувати модель можливого порушника. Ця модель може бути побудована з урахування різних критеріїв [5].

Модель порушника розробляється для того, щоб отримати відповіді на наступні питання [5]:

* від кого захищати інформацію?
* яка мета порушника?
* якими знаннями володіє порушник?
* які повноваження в системі має потенційний порушник?
* якими методами і засобами користується порушник?
* яка обізнаність порушника щодо об’єкта інформаційної діяльності і системи охорони [5]?

Інформація – відомості про об’єкти та явища навколишнього середовища, їхні параметри, властивості й стани, які зменшують наявну про них ступінь невизначеності, неповноти знань [5].

Теоретична інформатика розглядає інформацію як концептуально зв’язані між собою відомості, дані, поняття, що змінюють уявлення про явище або об’єкт навколишнього світу. Поряд з інформацією в інформатиці часто використовують поняття дані. Дані можуть розглядатися як ознаки або записані спостереження, які з будь яких причин не використовуються, а тільки зберігаються. У тому випадку, якщо з’являється можливість використати ці дані для визначення невизначеності будь-чого, дані перетворюються в інформацію. Тобто інформація – це дані, які використовуються [5].

Модель порушника представляє собою опис можливих дій порушника, який складається на основі аналізу типу зловмисника, рівня його повноважень, знань, теоретичних та практичних можливостей. Порушників прийнято поділяти на зовнішніх і внутрішніх. До внутрішніх належать співробітники, користувачі інформаційної системи, які можуть наносити шкоду інформаційним ресурсам як ненавмисно, так і навмисно; технічний персонал, який обслуговує будівлі і приміщення (електрики, сантехніки, прибиральниці тощо); персонал, який обслуговує технічні засоби (інженери, техніки). Зовнішні порушники - це сторонні особи, які знаходяться поза контрольованою зоною організації або не авторизовані для використання даної комп’ютерної системи. Це означає, що вони не мають в системі облікового запису і згідно системної політики безпеки взагалі не можуть працювати в даній системі. Приклад зовнішніх порушників: відвідувачі, які можуть завдати шкоди навмисно або через незнання існуючих обмежень; кваліфіковані хакери; особи, яких найняли конкуренти для отримання необхідної інформації; порушники пропускного режиму [5].

При розробці моделі порушника необхідно визначитись, що і у якій мірі має відображати отримана модель. Для цього необхідно визначитись з необхідним ступенем деталізації моделі порушника [5].

Можна запропонувати наступні ступені деталізації [5]:

* змістовна модель порушників - відображає причини й мотивацію дій порушників, переслідувані ними цілі і загальний характер дій у процесі підготовки і здійснення порушення інформаційної безпеки. Побудувавши змістовну модель, адміністратори безпеки можуть визначити мету порушника, його рівень знань, кваліфікацію, розташування та т.п [5].
* сценарії впливу порушників - визначають класифіковані типи порушень з конкретизацією алгоритмів і етапів, а також способи дії на кожному етапі. Розробивши сценарії впливу, адміністратори безпеки отримають можливу послідовність дій зловмисника для нанесення збитків інформаційним ресурсам [5].
* математична модель впливу порушників представляє собою формалізований опис сценаріїв у вигляді логіко-алгоритмічної послідовності дій порушників, кількісних значень, що параметрично характеризують результати дій, і функціональних (аналітичних, числових чи алгоритмічних) залежностей, які описують процеси взаємодії порушників з елементами об’єкту і системи охорони. Цей вид моделі слід використовувати для кількісних оцінок вразливості об’єкту і ефективності охорони [5].

Для того, щоб модель порушника найбільш точно і детально характеризувала порушників, алгоритм їх дій і давала кількісні оцінки вразливості об’єкту і ефективності охорони рекомендується розробляти комплексну модель з урахуванням усіх ступенів деталізації [5].

Способи класифікацій порушників

Під час побудови моделі порушника спочатку необхідно проаналізувати усіх користувачів системи, розподілити їх за категоріями та визначити найбільш критичні. Користувачі таких категорій будуть прийняті як можливі внутрішні порушники системи. Далі необхідно визначитись, які категорії відвідувачів можуть бути зовнішніми порушниками [5].

Усіх можливих порушників необхідно класифікувати за різними показниками для того, щоб надалі скласти модель порушника. Нижче наведені можливі види класифікацій [5]:

* Класифікація порушників інформаційної безпеки за метою порушення. Класифікація за метою порушення проводиться для визначення мотивів порушника. Дії порушника в залежності від мети можуть бути спрямовані як на інформацію, так і на матеріальні носії інформації. Знаючи мету порушника, адміністратори безпеки будуть орієнтуватись, на захист якого ресурсу необхідно приділити більше уваги першочергово [5].
* Класифікація порушників інформаційної безпеки за рівнем знань про автоматизовані системи. Кожен порушник має певний рівень кваліфікації та поінформованості відносно організації функціонування лабораторії зовнішніх та внутрішніх мереж інформаційного комп’ютерного комплексу. В залежності від рівня знань, якими володіє порушник, може бути нанесений певний рівень збитків інформаційним ресурсам організації. В класифікації враховуються знання можливого порушника та його практичні навички у роботі з комп’ютерними системами та інформаційними технологіями [5].
* Класифікація порушника за місцем дії. Ця класифікація проводиться для визначення розташування порушника відносно організації під час здійснення спроби несанкціонованого доступу до інформаційного ресурсу [5].
* Класифікація порушників за методами і способами, якими вони користуються. Порушник може отримати конфіденційну інформацію та інформацію з обмеженим доступом, користуючись при цьому різними методами та засобами. Порушення може бути скоєне або з використанням певних засобів для отримання інформації, або без них. Методи можуть бути різними, як дозволеними, так і забороненими. Дозволеним вважається отримання інформації без порушення прав власності. Як приклад можна привести використання методів соціальної інженерії [5].
* Класифікація порушників за рівнем можливостей, які надані їм засобами автоматизованої системи та обчислювальної техніки. Внутрішніх порушників можна класифікувати за наданим рівнем повноважень у системі. Адже чим більше повноважень, там більше можливостей доступу до інформації з обмеженим доступом [5].
* Класифікація порушників за мотивом порушень. Зловмисники можуть порушувати інформаційну безпеку з різних причин. Порушення можна розбити на дві групи - навмисні та ненавмисні. Особи, які ненавмисно наносять збитків інформаційним ресурсам, порушуючи конфіденційність, цілісність або доступність інформації, не складають плану дій, не мають мети та спеціальних методів та засобів реалізації запланованого порушення. Ненавмисні порушення частіше всього здійснюються в результаті недостатньої кваліфікації, неуважності персоналу. Порушники, які наносять збитків інформаційним ресурсам навмисно, мають певну мету, готують план реалізації атаки на інформаційний ресурс. Навмисні порушення інформаційної безпеки здійснюються для нанесення збитків організації (матеріальних чи моральних), для власного збагачення за рахунок отриманої інформації, а також для нейтралізації конкурентів [5].

## 1.5. Сучасні кіберзагрози інформаційному простору

Виділяють наступні сучасні кіберзагрози [7]:

* соціальна інженерія та фішинг;
* вірусне програмне забезпечення;
* використання неактуальних версій програмного забезпечення;
* інсайдерські загрози;
* відсутність політик і процедур щодо поводження з інформаційними ресурсами [7].

Соціальна інженерія і фішинг

Соціальна інженерія базується на експлуатації людських слабкостей. В результаті успішного психологічного підстроювання до жертви досвідчений зловмисник може виявити багато базових моментів в роботі організації для планування злому, викрадення інформації. Це і робота системи контролю фізичного доступу, робота охорони, графік роботи прибиральниць, місцезнаходження принтерів, сміттєвих кошиків, наявність шредерів і т. д. Підготовка до проникнення в інформаційну систему організації починається саме з такої роботи [7].

Фішингові атаки є продовженням соціальної інженерії. Багатьма фахівцями вони визнаються найбільш масовим і ефективним засобом злому і подальшого доступу до ресурсів підприємств і організацій всіх форм власності. За різними оцінками, до 90% всіх успішних кібератак відбуваються з використанням даного методу. Метод, по суті, дуже простий і являє собою розсилку підроблених листів електронної пошти, текст яких спонукає передбачувану жертву запустити вірусну програму, замасковану, наприклад, під офісний додаток, або перейти по посиланню на підроблений сайт, на якому вам запропонують ввести свій логін і пароль до пошти або інших ресурсів [7].

Вірусне програмне забезпечення

Це найважливіша зброю зловмисників, для захисту від якого необхідно мати актуальну версію антивірусного програмного забезпечення. Не існує антивірусного ПЗ, яке б захищало від всіх вірусів однаково добре. Як показує практика, дуже корисний обмін інформацією між фахівцями про початок атак і появі нових вірусів, а також використання служб з аналізу підозрілих файлів і посилань на предмет виявлення черв'яків, троянів і всіляких шкідливих програм [7].

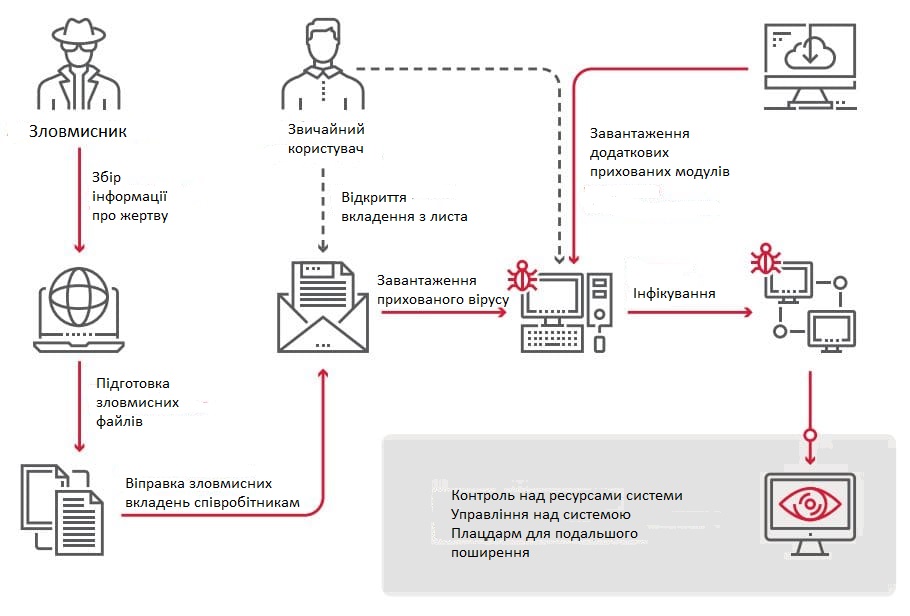


Рис. Етапи злому інформаційної системи

Неактуальні версії програмного забезпечення

По суті, ця вразливість інформаційної системи, коли Ви пропустите установку необхідних оновлень ПЗ. Регулярне оновлення ПЗ необхідно з багатьох причин, одна з яких - підвищення рівня безпеки та захист від нововиявлених вразливостей і загроз. Потрібно працювати з персоналом в цьому напрямку, робити регулярні розсилки і нагадування для співробітників, показувати на прикладах, до чого може привести використання неактуальною версії ПЗ [7].

Інсайдерські загрози



Рис. Механізм роботи кіберзагроз на інформаційні системи

Це велика група загроз, джерелом яких є власні співробітники. поширена в організаціях, де відсутній контроль за наданням прав доступу високого рівня, а також розмежування з прав доступу до інформаційних ресурсів. Дуже добре, якщо регулярно проводиться хоча б мінімальна оцінка лояльності співробітників при прийомі на роботу, в процесі роботи і при звільненні. Персонал - це завжди найслабша ланка в системі безпеки і з ним потрібно постійно працювати. Не дарма британські та американські фахівці з безпеки будують свою роботу саме з розвитку культури безпеки [7].

Відсутність політик і процедур щодо поводження з інформаційними активами.

Політики і процедури - це встановлені правила роботи в інформаційній системі організації, а також розподіл ролей і відповідальності. Без них настає хаос, можливі фінансові втрати через неефективне використання ресурсів, виникає підвищений ризик помилок персоналу, ненавмисна витік та втрата інформації. При їх наявності, навпаки, користувач може сам вирішувати багато питань і працювати в чітко окресленому полі. При успішно працюючій процедурі надання прав доступу виконується одна з базових правил інформаційної безпеки - доступ тільки до того, що необхідно по роботі [7].

При наявності правильно працюють політик безпеки не виникає питань, хто за що відповідає, які ресурси слід захищати в першу чергу, до кого звертатися в разі настання інцидентів кібербезпеки і скільки часу організація може працювати без серйозних втрат в разі тимчасової зупинки тих чи інших процесів. Все це можна розрахувати заздалегідь в ході виконання процедури оцінки ризиків та аналізу наслідків інцидентів для бізнесу [7].

За інноваційними аналітичними моделями DARPA кіберзагрози поділяються [7]:

* загрози підключення до каналів зв’язку;
* загрози наявності недокументованих можливостей;
* загрози зняття інформації в момент передачі по мережі;
* загрози підміни інформації в каналі управління;
* Загроза підключення до каналів зв’язку [7].

Цифрова обробка сигналів дає можливість копіювання («відгалуження») трафіку в межах комунікаційної мережі без яких би то не було демаскуючих ознак. Факт копіювання неможливо відстежити, він не викликає зміни в каналі зв’язку, ні спотворень, пов'язаних з задержкою передачі. Це є якісною відмінністю сучасних інформаційних систем і мереж [7].

Загрози наявності недокументованих можливостей.

Недокументовані можливості самих систем (в особливості IP-мереж) є ще однією загрозою для конфіденційності інформації в системах, які захищаються. Програмне забезпечення систем обробки інформації являє собою складний програмний комплекс, в т.ч. який реалізує стек протоколів TCP / IP, і може містити [7]:

* недокументовані можливості, внесені розробниками з метою тестування або на певних етапах розробки нових функціональних можливостей систем [7];
* помилки в реалізації, наприклад, що призводять до уразливостей класу «переповнення буфера», і дозволяють отримати повний контроль над програмним забезпеченням системи до її перезавантаження [7].

Загрози зняття інформації в момент передачі по мережі.

Різні варіанти реалізацій загроз прослуховування трафіку традиційні для комп'ютерних мереж, що використовують у своїй структурі широкомовні сегменти (Ethernet, в т.ч. комутований, радіо-Ethernet і т.п.), і створюють ще один рівень можливих атак на системи IP-мереж. При відсутності шифрування трафіку на мережевому або більш високих рівнях моделі OSI існує кілька варіантів порушення конфіденційності переданих повідомлень [7].

При отриманні зловмисником адміністративних прав на комутуюче або маршрутизуючого обладнання (наприклад, в результаті атаки на комп'ютер адміністратора або при перехопленні його пароля, що передається у відкритому вигляді) у нього з'являються набагато могутніші засоби перехоплення IP-трафіку [7].

Методика централізованого управління IP-пристроями та комп’ютерами містить ще один можливий шлях прозорого для абонентів перехоплення їх даних. У момент встановлення IP-з'єднання початковий обмін інформацією, що містить їх імена, технічні можливості систем і т.п., в т.ч. IP-адреси кінцевих пристроїв, йде між серверами IP-мереж. На цьому етапі можлива підміна (засобами атак мережевого рівня) інформації про один або обох IP-адреси з метою впровадження противника в ланцюжок передачі трафіку за принципом прозорого проксі-сервера [7].

Подібний клас атак залишається абсолютно непомітним на прикладному рівні, тому що користувачу зазвичай не видно мережеві координати віддаленого абоненту, а стек протоколів не здатний показати факт підміни, і може бути виявлений тільки за допомогою спеціалізованого моніторингу мережевого трафіку [7].

## 1.6. Аналіз кіберзагроз

Аналіз загроз інформаційній безпеці дозволяє виділити складові сучасних комп'ютерних загроз - їх джерела і рушійні сили, способи і наслідки реалізації. Аналіз виключно важливий для отримання всієї необхідної інформації про інформаційні загрози, визначення потенційної величини збитку, як матеріальної, так і нематеріальної, і вироблення адекватних заходів протидії [8].

При аналізі загроз інформаційній безпеці використовуються три основні методи [8]:

* пряма експертна оцінка;
* статистичний аналіз;
* факторний аналіз [8].

Розглянемо наведені методи докладніше [8]:

Пряма експертна оцінка. Метод експертних оцінок заснований на тому, що параметри загроз задаються експертами. Експерти визначають переліки параметрів, що характеризують загрози інформаційній безпеці, і дають суб'єктивні коефіцієнти важливості кожного параметра [8].

Статистичний аналіз - це аналіз інформаційних загроз на основі накопичених даних про інциденти інформаційної безпеки, зокрема, про частоту виникнення загроз певного типу, їх джерела та причини успіху чи неуспіху реалізації. Наприклад, знання частоти появи загрози дозволяє визначити ймовірність її виникнення за певний проміжок часу. Для ефективного застосування статистичного методу потрібна наявність досить великий за обсягом бази даних про інциденти. Потрібно відзначити ще одну вимогу: при використанні великого об’єму інформації необхідні інструменти узагальнення даних і виявлення в базі вже відомої та нової інформації [8].

Аналіз заснований на виявленні факторів, які з певною ймовірністю ведуть до реалізації загроз і тим або іншим негативним наслідкам. Такими факторами можуть бути наявність привабливих для кіберзлочинців інформаційних активів, уразливості інформаційної системи, високий рівень вірусної активності в зовнішньому середовищі і т.д. Оскільки на сучасні інформаційні системи впливають безліч чинників, зазвичай використовується багатофакторний аналіз [8].

При аналізі загроз інформаційній безпеці найбільш ефективно застосовувати комплекс різних аналітичних методів. Це значно підвищує точність оцінки [8].

Методів інтелектуального аналізу

Технології аналізу даних, що базуються на застосуванні класичних статистичних підходів, мають низку недоліків. Відповідні методи ґрунтуються на використанні усереднених показників, на підставі яких важко з’ясувати справжній стан справ у досліджуваній сфері. Методи математичної статистики виявилися корисними насамперед для перевірки заздалегідь сформульованих гіпотез та «грубого» розвідницького аналізу, що становить основу оперативної аналітичної обробки даних (OLAP). Окрім того, стандартні статистичні методи відкидають (нехтують) нетипові спостереження – так звані піки та сплески. Проте окремі нетипові значення можуть становити самостійний інтерес для дослідження, характеризуючи деякі виняткові, але важливі явища [8].

Ці недоліки статистичних методів спонукали до розвитку нових методів дослідження складних систем, які останнім часом все частіше застосовуються для вирішення практичних завдань – методів інтелектуального аналізу даних. Інтелектуальний аналіз даних (далі ІАД) – виявлення прихованих закономірностей або взаємозв’язків між змінними у великих масивах необроблених даних [8].

Сфера застосування ІАД нічим не обмежена – вона скрізь, де є якісь дані. Але насамперед методи ІАД сьогодні зацікавили комерційні підприємства, що розгортають свої проекти на основі інформаційних сховищ даних (Data Warehousing). ІАД являють собою велику цінність для керівників і аналітиків у їх повсякденній діяльності. Ділові люди усвідомили, що за допомогою методів ІАД вони можуть одержати відчутні переваги у конкурентній боротьбі. Досвід багатьох підприємств показує, що віддача від використання ІАД може сягати 1000 % [8].

В англомовній літературі замість терміна «інтелектуальний аналіз даних» зазвичай використовується термін Data Mining (дослівний переклад – «видобуток даних»). «Data Mining – це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності» [8].

В основу інтелектуального аналізу покладена концепція шаблонів (паттернів), що відбивають фрагменти багатоаспектних взаємин у даних. Ці шаблони являють собою закономірності, властиві підвибіркам даних, які можуть бути компактно виражені у зрозумілій людині формі. Пошук шаблонів проводиться методами, не обмеженими апріорними припущеннями про структуру вибірки, та видами розподілів значень аналізованих показників. Важливе положення інтелектуального аналізу – нетривіальність розшукуваних шаблонів. Це означає, що знайдені шаблони повинні відбивати неочевидні, несподівані регулярності в даних, що становлять так звані приховані знання. До суспільства прийшло розуміння, що сирі дані містять глибинний шар знань, за грамотного «розкопування» якого можуть бути виявлені справжні «самородки» [8].

Розглянемо основні задачі, які вирішуються методами Data Mining:

* класифікація – віднесення об'єктів (спостережень, подій) до одного з заздалегідь відомих класів [9];
* регресія (в тому числі задачі прогнозування) – встановлення залежності безперервних вихідних від вхідних змінних [9];
* кластеризація – угруповання об'єктів (спостережень, подій) на основі даних (властивостей), що описують сутність цих об'єктів, у кластери. Об'єкти всередині кластера повинні бути «схожими» один на одного і відрізнятися від об'єктів, що ввійшли в інші кластери. Чим більше схожі об'єкти всередині кластера і чим більше відмінностей між кластерами, тим точніша кластеризація [9];
* асоціація – виявлення закономірностей між пов'язаними подіями. Прикладом такої закономірності служить правило, яке вказує, що з події X слід подія Y. Такі правила називаються асоціативними. Вперше ця задача була запропонована для знаходження типових шаблонів покупок, що здійснюються в супермаркетах, тому іноді її ще називають аналізом ринкової корзини (market basket analysis) [9];
* послідовні шаблони – встановлення закономірностей між пов'язаними в часі подіями, тобто виявлення залежності, що якщо відбудеться подія X, то через заданий час відбудеться подія Y [9];
* аналіз відхилень – виявлення найбільш нехарактерних шаблонів.

Інтелектуальний аналіз даних може допомогти у вирішенні проблем інформаційної безпеки шляхом пошуку моделей, асоціацій, кореляцій, які приховані в інформації, що зберігається в базах даних або передаються по каналам зв’язку [10, 11, 12].

## 1.7. Система управління інформаційною безпекою

Система управління інформаційною безпекою (Information Security Management System, ISMS) – це частина загальної системи управління, що базується на аналізі ризиків і призначена для проектування, реалізації, контролю, супроводження та вдосконалення заходів у галузі інформаційної безпеки. Цю систему складають організаційні структури, політика, дії з планування, обов’язки, процедури, процеси і ресурси [6].

Найбільш значущою метою більшості систем інформаційної безпеки є захист бізнесу та знань компанії від знищення або витоку. Також однією з основних цілей системи інформаційної безпеки є гарантія майнових прав та інтересів клієнтів. У той же час заходи з інформаційної безпеки не повинні обмежувати або ускладнювати процеси обміну інформацією в компанії, оскільки це може поставити під загрозу розвиток організації [6].

Система управління інформаційною безпекою повинна забезпечувати гарантію досягнення таких цілей як забезпечення конфіденційності критичної інформації, забезпечення неможливості несанкціонованого доступу до критичної інформації, цілісності інформації та пов’язаних з нею процесів (створення, введення, обробки і виведення) і ряду інших цілей [6].

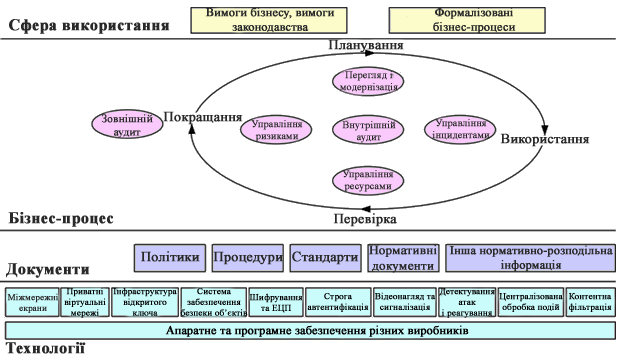


Рис. 1.1. Система управління інформаційною безпекою.

Досягнення заданих цілей можливо у ході вирішення таких основних завдань, як визначення відповідальних за інформаційну безпеку, розробка спектра ризиків інформаційної безпеки та проведення їх експертних оцінок, розробка політик і правил доступу до інформаційних ресурсів, розробка системи управління ризиками інформаційної безпеки, у тому числі методи їх оцінки, контролю інформаційної безпеки на підприємстві. Слід зазначити, що тут перераховано не повний список [6].

Побудова СУІБ дозволяє чітко визначити, як взаємопов’язані процеси та підсистеми ІБ, хто за них відповідає, які фінансові та трудові ресурси необхідні для їх ефективного функціонування, і т.д [6].

Основні функції системи управління інформаційною безпекою:

* виявлення та аналіз ризиків інформаційної безпеки;
* планування та практична реалізація процесів, спрямованих на мінімізацію ризиків ІБ;
* контроль цих процесів;
* внесення в процеси мінімізації інформаційних ризиків необхідних коригувань [6].

Якісне управління інформаційною безпекою базується на наступних принципах [6]:

* комплексний підхід – управління ІБ має бути всеосяжним, охоплювати всі компоненти ІС і враховувати всі актуальні ризикоутворюючі фактори, що діють в інформаційній системі підприємства та за її межами;
* узгодженість з бізнес-задачами і стратегією підприємства;
* високий рівень керованості;
* адекватність інформації, яка використовується і генерується;
* ефективність – оптимальний баланс між можливостями, продуктивністю і витратами СУІБ;
* безперервність управління;
* процесний підхід – зв’язування процесів управління в замкнутий цикл планування, впровадження, перевірки, аудиту та коригування, і підтримка нерозривного зв’язку між етапами [6].

Одним з ключових чинників успішності системи управління інформаційною безпекою підприємства – це побудова її на базі міжнародних стандартів ISO/IEC 31801.

Міжнародний стандарт ISO 31801 надає інструмент для розробки, впровадження, супроводу, моніторингу, підтримки та вдосконалення добре документованої системи управління інформаційною безпекою в контексті розгляду бізнес ризиків. СУІБ забезпечує вибір адекватних і пропорційних методів і засобів контролю та захисту інформації і, тим самим, довіру зацікавлених сторін [6].

Проте слід брати до уваги й інші стандарти в сфері інформаційної безпеки. На даний момент у світовій практиці використовується велика кількість стандартів, методик та інших документів, що регламентують процеси управління інформаційною безпекою, наприклад ISM3, COBIT, ITIL / ITSM, BSI-100-3, ISO13335-4, CRAMM, ISO15408. Але варто відмітити, що всі вони сумісні з ISO 31801, а також подобні до нього [6].

## 1.8. Висновок до розділу

В даному розділі на основі Закону України «Про основні засади забезпечення кібербезпеки України» розглянуто перелік загроз і рекомендацій щодо їх усунення, на які не завжди керівники і власники бізнесу, керівники державних організацій звертають належну увагу. Хоча важливість цих загроз, від яких захищаються далеко не завжди технічними заходами, складно переоцінити.

Багато хто намагається сховатися за технологією, ігноруючи роботу з людьми і процедури. Потрібно приділяти достатню увагу всієї тріаді безпеки: люди, процедури, технології. Кібергігіена вже давно широко застосовується в ефективно працюючих компаніях всього світу. Знаючи свої ризики, необхідно працюйте з людьми, впроваджувати технології та регулювати правила роботи з політиками і процедурами.

При досліджені різних видів кіберзагроз основною і часто надважливою складовою є канал зв’язку жертви в інформаційній системі і системи управління зловмисника, або інсайдера з його кураторами. Часто даний канал зв’язку є прихований. Тому дослідження даних каналів зв’язку є найважливішим з етапів дослідження кіберзагроз в цілому.

# Розділ 3. Аналіз прихованої інформації, машинне навчання

## 3.1. Класифікація методів стеганографії

На сьогодні вченими розроблені і випробувані різні алгоритми і методи стеганографії, ми відзначимо наступні:

* LSB-стеганографія (повідомлення ховається в молодших бітах (можливе використання одного або декількох молодших біт) контейнера. Чим менше біт задіяно, тим менше артефактів отримує оригінальний контейнер після впровадження [13].
* Метод, заснований на приховуванні даних в коефіцієнтах дискретного косинусного перетворення (далі ДКП) - різновид попереднього методу, яка активно використовується, наприклад, при впровадженні повідомлення в контейнер формату JPEG. За інших рівних, такий контейнер має дещо меншу ємність ніж в попередньому методі, в тому числі за рахунок того, що коефіцієнти «0» і «1» залишаються незмінними - впровадження повідомлення в них неможливо [13].
* Метод приховування інформації за допомогою молодших біт палітри- цей метод по суті є варіантом загального методу LSB, але інформація вбудовується не в найменш значущі біти контейнера, а в найменш значущі біти палітри, очевидний недолік такого методу - низька ємність контейнера [13]..
* Метод приховування інформації в службових полях формату - досить простий метод, заснований на використанні службових полів заголовка контейнера для зберігання повідомлення. Очевидні мінуси - низька ємність контейнера і можливість виявлення впроваджених даних за допомогою звичайних програм для перегляду зображення (які іноді дозволяють бачити вміст службових полів) [13].
* Метод вбудовування повідомлення - полягає в тому, що повідомлення вбудовується в контейнер, потім за допомогою схеми, відомої обом сторонам, витягується. Можна вбудувати кілька повідомлень в один контейнер, за умови, що способи їх впровадження ортогональні[13].

Широкосмугові методи, які поділяються на:

* Метод псевдослучайной послідовності; використовується секретний сигнал, який моделюється псевдовипадковим сигналом [13].
* Метод стрибаючих частот: частота несучого сигналу змінюється за певним псевдослучайному закону [13].
* Метод накладення - по суті не є справжньою стеганографії, заснований на тому, що деякі формати містять в заголовку розмір даних, або ж обробник цих форматів буде читати файл до маркера кінця даних. Прикладом такого методу є добре відомий метод «rar-jpeg», який заснований на конкатенації графічного файлу в форматі JREG і RAR-архіву. ПО для перегляду JPEG буде зчитувати інформацію до кордону, зазначеної в заголовку файлу, а RAR-архіватор відкине всі, що знаходиться до сигнатури «RAR!», Яка позначає початок архіву. Таким чином, якщо такий файл відкрити у вікні перегляду графічних файлів - ми побачимо картинку, а якщо в RAR-архіватор - вміст RAR-архіву. Очевидні мінуси такого підходу полягають в тому, що оверлей, доданий до контейнера, легко виділяємо при візуальному дослідженні такого файлу [13].

## 3.3. Класифікація методів стегоаналізу

Як і будь-яке складне явище, методи стегоаналізу можна класифікувати за багатьма критеріями.

За широтою аналізу контейнерів, в яких ховаються дані, методи стегоаналізу діляться на [13]:

* методи, призначені для виявлення даних, прихованих певним алгоритмом [13];
* методи "сліпого" розпізнавання [13].

Звичайно ж, вузькоспрямовані методи, тобто орієнтовані на певний алгоритм, мають кращі показники знаходження прихованих даних. Їх недоліком є ​​те, що в разі невеликої зміни алгоритму приховування, шукані дані вже можуть не виявлятися. Перевагою "сліпих" або ж універсальних методів є широкий спектр стеганографічних алгоритмів приховування, над якими вони можуть працювати. Разом з тим ці методи потребують "навчання", від якості якого залежать їх стегоаналітичні можливості [13].

За впливом на файл-контейнер, який містить приховані дані, стегоаналітичні методи діляться на [13]:

* методи пасивного стегоаналізу, які визначають наявність/відсутність прихованих даних в стегоконтейнер, або що визначають алгоритм, за яким відбувалося вбудовування [13].
* методи активного стегоаналізу, які визначають довжину вбудованого документа, його розташування, деякі параметри алгоритму вбудовування, а також вилучають приховану інформацію [13].
* По об'єкту пошуку в стегоконтейнер стегоаналітичні методи можна розділити на такі типи [13]:
* сигнатурні методи, побудовані на пошуку в стеганограмах так званих "відбитків пальців" (fingerprints), - фрагментів коду, які залишають після своєї роботи стеганографічні програми [13];
* імовірнісні методи, які засновані на аналізі імовірнісних показників, характерних для стегоповідомлень [13].

До методів сигнатурного типу відносяться:

* візуальна атака на стегосистеми [13];
* аналіз відповідності формату даних (пошук зайвих включень) [13].

Так в певних видах зображень виявляються досить тісні зв'язки всіх біт зображення. Використовуючи для таких зображень програми-фільтри, можна побачити вбудовану інформацію, яка порушує наявні залежності. Візуальна атака повністю ґрунтується на здатності зорової системи людини аналізувати зорові образи і виявляти істотні розбіжності в порівнюваних зображеннях [13].

Прикладами "зайвої" інформації є додаткові байти інформації, які дописують в кінець графічного файлу, або дублювання квітів, в палітрі [13].

До ймовірнісних методів можна віднести:

* аналіз статистики Хі-квадрат;
* метод дослідження статистичних моделей вищого порядку.

Ідея статистичного стегоаналізу полягає в порівнянні теоретичного розподілу найменш значущих біт малюнка з фактичним розподілом цих же біт в цьому ж малюнку. При цьому міра подібності теоретичного і фактичного розподілу є мірою достовірності вбудовування прихованої інформації. Статистичні моделі вищого порядку ґрунтуються, наприклад, на вейвлет-декомпозиції зображень. За результатами декомпозиції будується "характеристичний" вектор, який використовується для диференціювання зображень з прихованими даними і «чистих». Найкращі результати модель дає після "тренування" на деякому наборі порожніх контейнерів і відповідних їм заповнених контейнерів [13].

За формою подання досліджуваного зображення, методи стегоаналізу можна розділити на:

* методи, які аналізують зображення безпосередньо, тобто в просторовій формі подання [13];
* методи, які аналізують частотні форми представлення зображення, тобто після переведення його до частотної формі за допомогою дискретного перетворення косинуса, або вейвлет-перетворення [13].

По виду математичних інструментів, які використовуються для визначення факту вбудовування, розрізняють:

* методи, які використовують статистичні критерії згоди (Хі-квадрат);
* методи, які застосовують заходи подібності цифровий сукупності;
* метод визначення JPEG- сумісності [13].

За аналогією з криптографією можливі інформаційно-теоретичний та теоретико-складні підходи до стегоаналізу. Перший підхід має на увазі необмежені обчислювальні і тимчасові витрати для приховує і для аналітика. При другому підході враховується обмеженість ресурсів, і мова йде про умовно стійкою стегосистеми. На практиці все стегосистеми можна вважати умовно стійкими, що як раз і відкриває можливість для проведення стегоаналізу [13].

Більшість методів стегоаналізу засноване на виявленні відхилення спостерігається мультимедійної інформації від його очікуваної моделі. Клас статистичних методів стегоаналізу використовує безліч статистичних характеристик, таких як: оцінка ентропії, коефіцієнти кореляції, ймовірність появи і залежності між елементами послідовностей, умовні розподілу, розрізнення розподілів за критерієм Хі-квадрат і багато інших. Найпростіші тести оцінюють кореляційні залежності елементів контейнерів, в які можуть впроваджуватися приховувані повідомлення [13].

Загальний недолік статистичних методів стегоаналізу полягає в тому, що побудова математичної моделі "природного зображення", або "природного аудіо сигналу", є виключно складним і досі не вирішеним завданням. Можна сказати, що між тим, хто приховує інформацію і стегоаналітиком відбувається змагання: хто з них володіє кращою моделлю контейнера, той і переможе в протиборстві [13].

Відомі методи стегоаналізу, спрямовані проти конкретних стегоалгоритмів або працюють для конкретних форматів даних. Наприклад, методи стегоаналізу зображень, в яких проводиться впровадження в найменший значущий біт, досліджують розподіл цих бітів, а також кореляційні зв'язки між ними і більш значущими бітами [13].

## 3.3.1. Стеганоаналіз методом критерію правдоподібності Хі-квадрат

Ступінь відмінності між імовірнісними розподілами елементів природних контейнерів і отриманих з них заповнених контейнерів може бути використана для оцінки ймовірності існування стегоканалу. Дану ймовірність зручно визначити з використанням критерію згоди Хі-квадрат . За критерієм Хі-квадрат порівнюється, наскільки розподіл досліджуваної послідовності близько до характерному для стегограм розподілу. У досліджуваної послідовності підраховується скільки разів її елемент прийняв значення, де всього *k* елементів. Наприклад, в гістограмі лівих і правих відрахунків кольору в лівій частині рис. відрахунок кольору «000» з'явився 3 рази (), а відрахунок «001» - 5 разів (). При встановленні чергових бітів приховуваного повідомлення в НЗБ цієї пари відрахунок кольору «000» повинен з'являтися в середньому раз [13]:

|  |  |
| --- | --- |
|  | [13] |

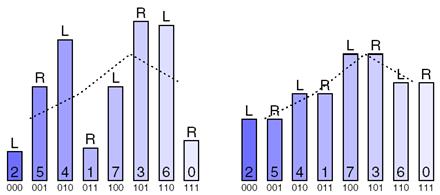


Рис. 3.1. Гістограма частот появлення лівих і правих номерів кольору, зліва – порожній контейнер, справа – заповнений.

Знаючи загальне число n появи всіх елементів досліджуваної послідовності, легко підрахувати очікувану ймовірність появи цих елементів в заповнених контейнерів за правилом [13]:

|  |  |
| --- | --- |
|  | [13] |

Відповідно, для досліджуваної послідовності ймовірності рівні [13]:

|  |  |
| --- | --- |
|  | [13] |

Величина Хі-квадрат для порівнюваних розподілів досліджуваної послідовності і очікуваного розподілу заповненого контейнеру дорівнює [13]:

|  |  |
| --- | --- |
|  | [13] |

де *v* є число ступенів свободи. Число ступенів свободи дорівнює числу k мінус число незалежних умов, накладених на ймовірності . Накладемо одну умову виду [13]:

|  |  |
| --- | --- |
|  | [13] |

Імовірність *p* того, що два розподілу однакові, визначається як, де Г є гамма-функція Ейлера [13].

|  |  |
| --- | --- |
|  | [13] |

Чим більше значення *p*, тим вище ймовірність вбудовування прихованої інформації в досліджувану послідовність [13].

## 3.3. Машинне навчання

### 3.3.1. Поняття машинного навчання

Машинне навчання являє собою підрозділ штучного інтелекту, що стоїть на стику таких дисциплін, як математика, статистика, теорія ймовірностей, теорія графів і вивчає алгоритми, які здатні самостійно навчатися на основі досвіду [9,14, 15, 16].

### 3.3.3. Види машинного навчання

Існує кілька основних способів машинного навчання:

* Навчання з вчителем;
* Навчання без вчителя;
* Часткове навчання;
* Навчання з підкріпленням;

При навчанні з учителем для кожного прецеденту задана пара «ситуація, рішення». Завдання такого навчання полягає в пошуку залежно прийнятого рішення від заданої ситуації і побудові алгоритму, здатного прийняти на вхід опис ситуації, а на виході передбачити для неї рішення [14].

При навчанні без учителя на вхід подаються тільки описи об'єктів без прийнятого рішення по цій ситуації, а завдання полягає в пошуку залежностей між представленими об'єктами [14].

Часткове навчання є проміжною ланкою між навчанням з учителем і без учителя, так як кожен прецедент задається парою «ситуація, рішення», однак відповіді відомі лише для частини цих ситуацій [14].

При навчанні з підкріпленням не існує «правильних відповідей» для кожної ситуації, алгоритм шукає оптимальну стратегію поведінки, спираючись на реакцію зовнішнього середовища [14].

### 3.3.3. Алгоритми машинного навчання

Існують різні алгоритми машинного навчання, на основі яких будується модель системи. Багато в чому вибір відповідного алгоритму залежить від характеристик набору даних, таких як обсяг, структура і якість. Також на вибір алгоритму впливає бажаний результат (двокласова або багатокласова класифікація, регресія або фільтрація викидів), необхідна точність передбачення і час, необхідний для навчання моделі [14].

Як правило, для передбачення значень змінної використовується регресійний аналіз. Його мета - розробити статистичну модель, що дозволяє прогнозувати значення залежної змінної, або відгуку, за значеннями, принаймні однієї, незалежної, або пояснює, змінної [14].

### 3.3.4. Види регресійних моделей

У замітці Подання числових даних у виді таблиць і діаграм для ілюстрації залежності між змінними *X* і *Y* використовувалася діаграма розкиду. На ній значення змінної *X* відкладалися по горизонтальній осі, а значення змінної *Y* - по вертикальній. Залежність між двома змінними може бути різною: від найпростішої до вкрай складною. Приклад найпростішої (лінійної) залежно показаний на рис [17, 18, 19].

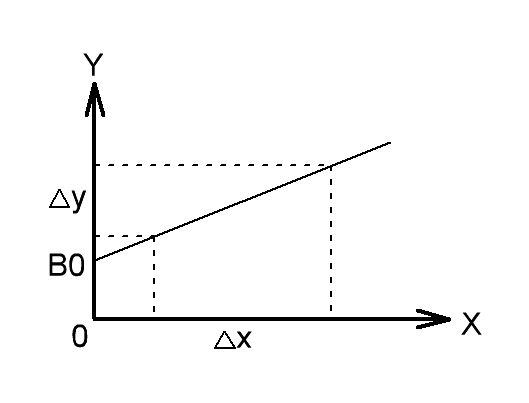


Рис. 3.3. Позитивна лінійна залежність

Проста лінійна регресія:

|  |  |
| --- | --- |
|  | [17] |

де - зсув (довжина відрізка, що відсікається на координатної осі прямої *Y* ), - нахил прямої *Y* , - випадкова помилка змінної *Y* в *i*-му спостереженні.

У цій моделі нахил являє собою кількість одиниць вимірювання змінної *Y* , що припадають на одну одиницю вимірювання змінної *X* . Ця величина характеризує середню величину зміни змінної *Y* (позитивного або негативного) на заданому відрізку осі *X* . Зрушення являє собою середнє значення змінної *Y* , коли змінна *X* дорівнює 0. Останній компонент моделі є випадковою помилкою змінної *Y* в *i*-му спостереженні. Вибір підходящої математичної моделі залежить від розподілу значень змінних *X* і *Y* на діаграмі розкиду. Різні види залежно змінних показані на рис. 3.3 [17, 18, 19].

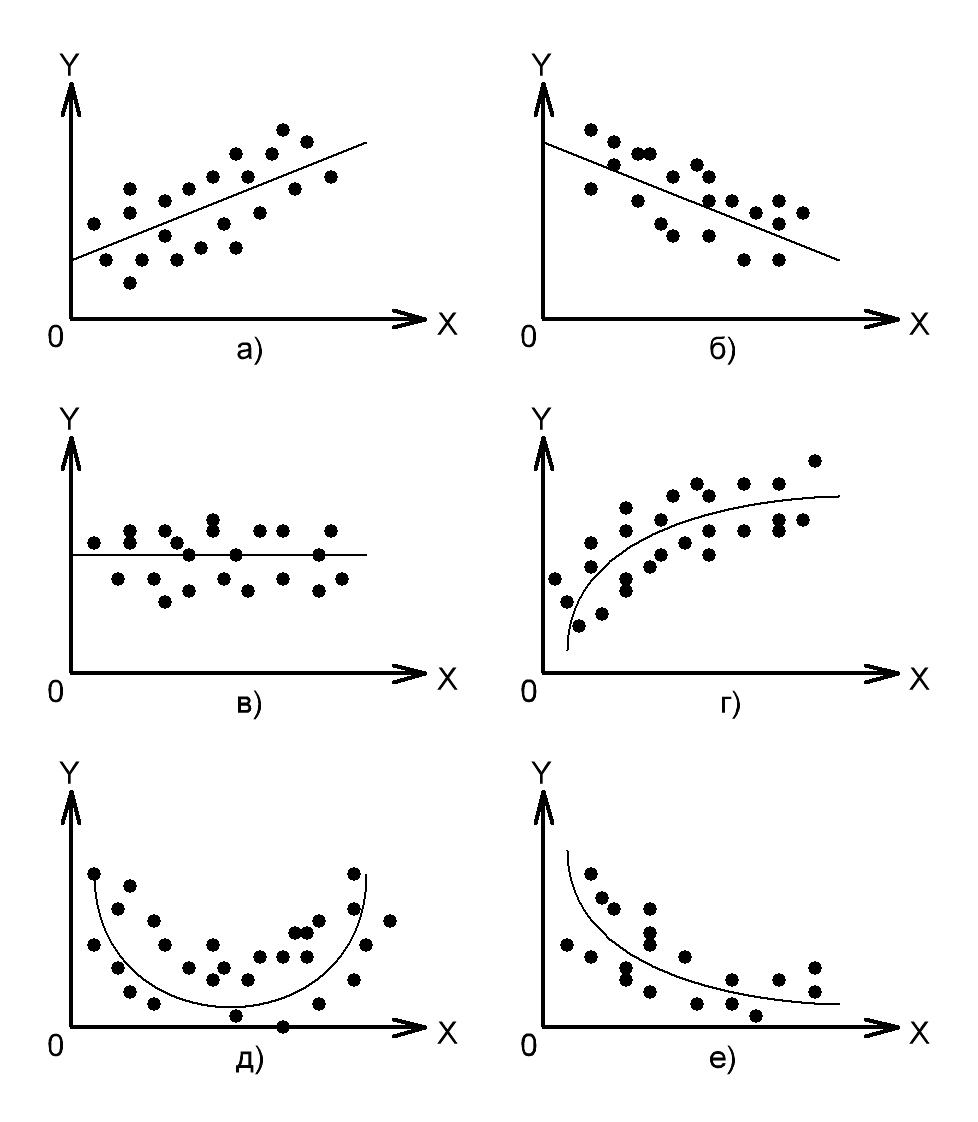


Рис. 3.3. Діаграми розкиду, що ілюструють різні види залежностей

На рис. а значення змінної *Y* майже лінійно зростають зі збільшенням змінної *X* . На рис. б) є прикладом негативної лінійної залежності. Якщо змінна *X* зростає, змінна *Y* в цілому зменшується. На рис. в) показаний набір даних, в якому змінні *X* і *Y* практично не залежать один від одного. Кожному значенню змінної *X* відповідають як більші, так і менші значення змінної *Y* . Дані, наведені на рис. г , демонструють криволінійну залежність між змінними *X* і *Y* . Значення змінної *Y* зростають при збільшенні змінної *X* , однак швидкість росту після певних значень змінної *X* падає. Рис. д) демонструє параболічну U-подібну форму залежності між змінними *X* і *Y* . У міру збільшення значень змінної *X* значення змінної *Y* спочатку зменшуються, а потім зростають. На панелі Е показана експоненціальна залежність між змінними *X* і *Y* . У цьому випадку змінна *Y* спочатку дуже швидко зменшується при зростанні змінної *X*, однак швидкість цього зменшення поступово падає [17, 18, 19].

Ми коротко розглянули основні моделі, які дозволяють формалізувати залежності між двома змінними. Незважаючи на те що діаграма розкиду надзвичайно корисна при виборі математичної моделі залежності, існують більш складні і точні статистичні процедури, що дозволяють описати відносини між змінними. Надалі ми будемо розглядати лише лінійну залежність [17, 18, 19].

### 3.3.5. Метод найменших квадратів

Дані, представлені на рис. а, отримані для випадкової вибірки. Якщо вірні деякі припущення (про це трохи пізніше), в якості оцінки параметрів генеральної сукупністю ( і ) можна використовувати зрушення і нахил прямої *Y* . Таким чином, рівняння простої лінійної регресії приймає наступний вигляд [14, 15, 16]:

|  |  |
| --- | --- |
|  | [14, 15] |

де - передбачене значення змінної *Y* для *i*-гo спостереження, - значення змінної *X* в *i*-му спостереженні.

Для того щоб спрогнозувати значення змінної *Y*, в рівнянні (3) необхідно визначити два коефіцієнта регресії - зрушення і нахил прямий *Y*. Обчисливши ці параметри, проведемо пряму на діаграмі розкиду. Потім дослідник може візуально оцінити, наскільки близька регресійна пряма до точок спостереження. Проста лінійна регресія дозволяє знайти пряму лінію, максимально наближену до точок спостереження. Критерії відповідності можна задати різними способами. Можливо, простіше за все мінімізувати різниці між фактичними значеннями , і передбаченими значеннями . Однак, оскільки ті різниці можуть бути як позитивними, так і негативними, слід мінімізувати суму їх квадратів [14, 15, 16].

Оскільки , сума квадратів приймає наступний вигляд:

|  |  |
| --- | --- |
|  | [14] |

Параметри і невідомі. Таким чином, сума квадратів різниць є функцією, яка залежить від зсуву і нахилу вибірки Y . Для того щоб знайти значення параметрів і , мінімізують суму квадратів різниць, застосовується метод найменших квадратів. При будь-яких інших значеннях зсуву і нахилу сума квадратів різниць між фактичними значеннями змінної і її спостерігаючим значенням лише збільшиться [14, 15, 16].

### 3.3.6. Інтерполяція і екстраполяція.

Застосовуючи регресійну модель для прогнозування, необхідно враховувати лише допустимі значення незалежної змінної. В цей діапазон входять всі значення змінної *X*, починаючи з мінімальної і закінчуючи максимальної. Таким чином, прогнозуючи значення змінної *Y* при конкретному значенні змінної *X* , дослідник виконує інтерполяцію між значеннями змінної *X* в діапазоні можливих значень. Однак екстраполяція значень за межі цього інтервалу не завжди релевантна. Будь-яка спроба екстраполяції означає, що ми припускаємо, ніби лінійна регресія зберігає свій характер за межами допустимого діапазону [17, 18, 19, 20, 21].

### 3.3.7. Обчислення сум квадратів.

Для того щоб передбачити значення залежної змінної за значеннями незалежної змінної в рамках обраної статистичної моделі, необхідно оцінити мінливість. Існує кілька способів оцінки мінливості. Перший спосіб використовує загальну суму квадратів (total sum of squares - SST), що дозволяє оцінити розкид значень *Y*i навколо середнього значення. У регресійному аналізі повна варіація, що представляє собою повну суму квадратів, розділяється на зрозумілу варіацію, або суму квадратів регресії (regression sum of squares - SSR), і незрозумілу варіацію, або суму квадратів помилок (error sum of squares - SSE). Зрозуміла варіація характеризує взаємозв'язок між змінними *X* і *Y*, а незрозуміла залежить від інших факторів [22, 23, 24, 25, 26].

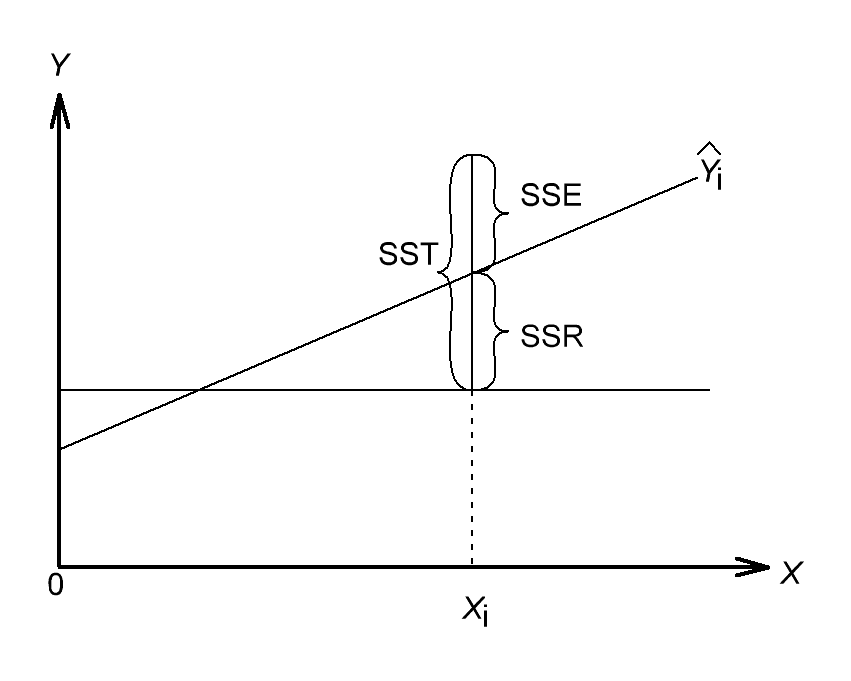


Рис. 3.4. Оцінки змінності в моделі регресії

Сума квадратів регресії (SSR) являє собою суму квадратів різниць між i (передбаченим значенням змінної *Y*) і (середнім значенням змінної *Y*). Сума квадратів помилок (SSE) є частиною варіації змінної Y , яку неможливо описати за допомогою регресійної моделі. Ця величина залежить від різниць між тим що спостерігається і прогнозованим значеннями [22, 23, 24, 25, 26].

Повна сума квадратів (SST) дорівнює сумі квадратів регресії плюс сума квадратів помилок:

|  |  |
| --- | --- |
| *SST = SSR + SSE* | [30] |

Повна сума квадратів (SST) дорівнює сумі квадратів різниць між тим що спостерігається значеннями змінної *Y* і її середнім значенням [30, 31, 33]:

|  |  |
| --- | --- |
|  | [30] |

Сума квадратів регресії (SSR) дорівнює сумі квадратів різниць між передбаченими значеннями змінної *Y* і її середнім значенням:

|  |  |
| --- | --- |
|  | [30] |

Сума квадратів помилок (SSE) дорівнює сумі квадратів різниць між тим що спостерігається і передбаченими значеннями змінної Y:

|  |  |
| --- | --- |
|  |  |

**3.3.8. Коефіцієнт змішаної кореляції.**

Величини SSR , SSE і SST не мають очевидної інтерпретації. Однак відношення суми квадратів регресії ( SSR ) до повної суми квадратів ( SST ) являє собою оцінку корисності регресійного рівняння. Це відношення називається коефіцієнтом змішаної кореляції :

|  |  |
| --- | --- |
|  |  |

Коефіцієнт змішаної кореляції оцінює частку варіації змінної *Y* , яка пояснюється незалежної змінної *X* на регресійній моделі.

### 3.3.9. Середньоквадратична помилка оцінки

Хоча метод найменших квадратів дозволяє обчислити лінію, що мінімізує відхилення від спостережуваних значень, наявність суми квадратів помилок (SSE) свідчить про те, що лінійна регресія не дає абсолютної точності прогнозу, якщо, звичайно, точки спостереження не лежать на регресійній прямій. Проте очікувати цього так само неприродньо, як припускати, що всі вибіркові значення точно рівні їх середньому арифметичному. Отже, необхідна статистика, яка дозволила б оцінити відхилення передбачених значень змінної *Y* від її реальних значень, аналогічно тому, як стандартне відхилення, введене раніше, дозволяє оцінити розкид даних навколо їх середньої величини. Стандартне відхилення спостережуваних значень змінної *Y* від її регресійної прямої називається середньоквадратичної помилкою оцінки [22, 23, 24, 25, 26].

Середньоквадратична помилка оцінки:

|  |  |
| --- | --- |
|  |  |

де - фактичне значення змінної *Y* при заданому значенні , - передбачене значення змінної *Y* при заданому значенні , - сума квадратів помилок.

Ефективне значення помилка оцінки характеризує відхилення реальних даних від лінії регресії. Вона вимірюється в тих же одиницях, що і змінна Y . За змістом середньоквадратична помилка дуже схожа на стандартне відхилення. У той час як стандартне відхилення характеризує розкид даних навколо їх середнього значення, середньоквадратична помилка дозволяє оцінити розкид точок спостереження навколо регресійної прямої. Середньоквадратична помилка оцінки дозволяє виявити статистично значущу залежність, існуючу між двома змінними, і спрогнозувати значення змінної *Y* [22, 23, 24, 25, 26].

### 3.3.10. Припущення

Обговорюючи методи перевірки гіпотез і дисперсійного аналізу, не раз підкреслювалось важливість умов, які повинні забезпечувати коректність зроблених висновків. Оскільки та регресійний, і дисперсійний аналізи використають лінійну модель, умови їх застосування приблизно однакові [27,28,29, 30, 31, 32]:

* Помилка повинна мати нормальний розподіл.
* Варіація даних навколо лінії регресії повинна бути постійною.
* Помилки повинні бути незалежними.

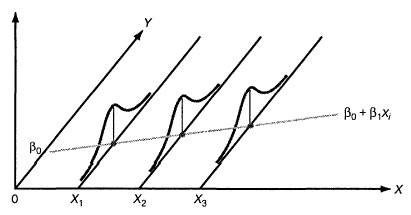


Рис. 3.5. Припущення про нормальний розподіл помилок

Перше припущення, про нормальний розподіл помилок, вимагає, щоб при кожному значенні змінної *X* помилки лінійної регресії мали нормальний розподіл (рис. 3.5.). Як і t- і F-критерій дисперсійного аналізу, регресійний аналіз досить стійкий до порушень цієї умови. Якщо розподіл помилок щодо лінії регресії при кожному значенні *X* не дуже сильно відрізняється від нормального, висновки щодо лінії регресії і коефіцієнтів регресії змінюються незначно [27,28,29, 30, 31, 32].

Друга умова полягає в тому, що варіація даних навколо лінії регресії повинна бути постійною при будь-якому значенні змінної *X* . Це означає, що величина помилки як при малих, так і при великих значеннях змінної *X* повинна змінюватися в одному і тому ж інтервалі (рис. ). Це властивість дуже важлива для методу найменших квадратів, за допомогою якого визначаються коефіцієнти регресії. Якщо ця умова порушується, слід застосовувати або перетворення даних, або метод найменших квадратів з вагами [27,28,29, 30, 31, 32].

Третє припущення, про незалежність помилок, полягає в тому, що помилки регресії не повинні залежати від значення змінної *X*. Ця умова особливо важливо, якщо дані збираються протягом певного відрізку часу. У цих ситуаціях помилки, властиві конкретному відрізку часу, часто корелюються з помилками, характерними для попереднього періоду [27,28,29, 30, 31, 32].

### 3.3.11. Аналіз залишків

Розглянемо тепер аналіз помилок - графічний метод, що дозволяє оцінити точність регресійній моделі. Крім того, з його допомогою можна виявити потенційні порушення умов застосування регресивного аналізу [32 ,33,34, 35, 36, 37].

### 3.3.13. Оцінка придатності емпіричної моделі

Залишок, або оцінка помилки , являє собою різницю між спостережуваним () і передбаченим () значеннями залежної змінної при заданому значенні X i [32 ,33,34, 35, 36, 37].

|  |  |
| --- | --- |
|  |  |

Для оцінки придатності емпіричної моделі регресії залишки відкладаються по вертикальній осі, а значення - по горизонтальній. Якщо емпірична модель придатна, графік не повинен мати яскраво вираженою закономірності. Якщо ж модель регресії не придатна, то на графіку проявиться залежність між значеннями і залишками [32 ,33,34, 35, 36, 37].

Розглянемо приклади (рис. 8). Рис. а) ілюструє зростання змінної *Y* при збільшенні змінної *X* . Однак залежність між цими змінними носить нелінійний характер, оскільки швидкість зростання змінної *Y* падає при збільшенні змінної *X* . Таким чином, для апроксимації залежності між цими змінними краще підійде квадратична модель. Особливо яскраво квадратична залежність між величинами *X* і *e* проявляється на панелі Б. Графічне зображення залишків дозволяє відфільтрувати або видалити лінійну залежність між змінними *X* і *Y* і виявити недостатню точність моделі простої лінійної регресії. Отже, в даній ситуації замість простої лінійної моделі повинна застосовуватися квадратична модель, що володіє більш високою точністю [32 ,33,34, 35, 36, 37].

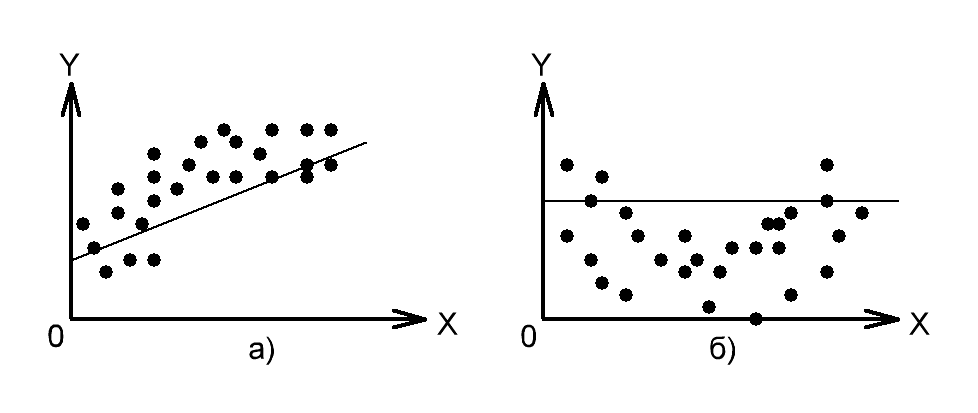


Рис. 3.6. Дослідження емпіричної моделі простої лінійної регресії

Перевірка умов. Графік залишків дозволяє оцінити варіації помилок. Розглянемо гіпотетичну ситуацію, в якій умова - варіації помилок при різних значеннях *Хi* приблизно однакові, не виконується (рис. 11). На цьому малюнку зображено ефект віяла: при зростанні значень *Хi* помилки збільшуються. Таким чином, мінливість значень *Yi* при різних значеннях *Хi* є непостійною [32 ,33,34, 35, 36, 37].

Незалежність. Припущення про незалежність помилок також перевіряється за допомогою графіка залишків. Дані, зібрані протягом деякого періоду часу, іноді демонструють ефект автокореляції між послідовними спостереженнями. У таких ситуаціях залишки залежать від значень попередніх залишків. Подібна зв'язок між залишками порушує припущення про незалежність помилок. Ефект автокореляції добре виявляється на графіку. Крім того, його можна виміряти за допомогою процедури Дурбіна-Уотсона. Якщо дані *X* і *Y* збиралися протягом одного і того ж періоду часу, гіпотезу про їх незалежності перевіряти не має сенсу [32 ,33,34, 35, 36, 37].

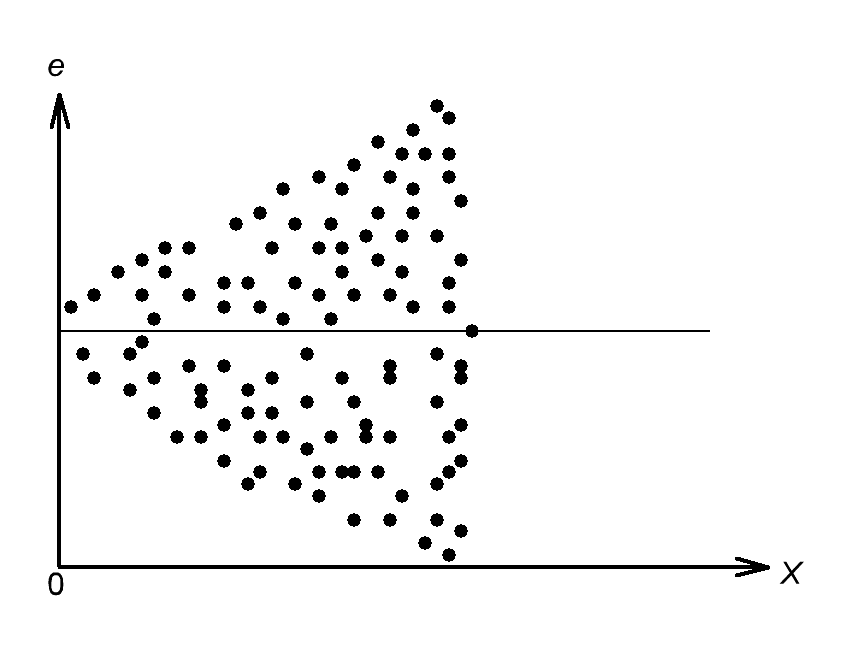


Рис. 3.7. Приклад порушення умови незалежності варіацій помилок від *Xi*

### 3.3.14. Перевірка гіпотез про нахил і коефіцієнті кореляції

Вище регресія застосовувалася виключно для прогнозування. Для визначення коефіцієнтів регресії і передбачення значення змінної *Y* при заданій величині змінної *X* використовувався метод найменших квадратів. Крім того, ми розглянули середньоквадратичну помилку оцінки та коефіцієнт змішаної кореляції. Якщо аналіз залишків підтверджує, що умови застосовності методу найменших квадратів не порушуються, і модель простої лінійної регресії є адекватною, на основі вибіркових даних можна стверджувати, що між змінними в генеральної сукупності існує лінійна залежність [32 ,33,34, 35, 36, 37].

Застосування t -критерію для нахилу

Перевіряючи, чи рівний нахил генеральної сукупності *β1* нулю, можна визначити, чи існує статистично значуща залежність між змінними *X* і *Y*. Якщо ця гіпотеза відхиляється, можна стверджувати, що між змінними *X* і *Y* існує лінійна залежність. Нульова і альтернативна гіпотези формуються наступним чином: (немає лінійної залежності), (є лінійна залежність). За визначенням t-статистика дорівнює різниці між вибірковим нахилом і гіпотетичним значенням нахилу генеральної сукупності, поділеній на середньоквадратичну помилку оцінки нахилу [32 ,33,34, 35, 36, 37]:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

де - нахил прямої регресії за вибірковими даними, - гіпотетичний нахил прямої генеральної сукупності, а тестова статистика *t* має t -розподіл з n - 3 ступенями свободи [32 ,33,34, 35, 36, 37].

Застосування F-критерію для нахилу.

Альтернативним підходом до перевірки гіпотез про нахил простої лінійної регресії є використання F-критерію. Нагадаємо, що F-критерій застосовується для перевірки відносини між двома дисперсіями. Під час перевірки гіпотези про нахилі мірою випадкових помилок є дисперсія помилки (суму квадратів помилок, поділена на кількість ступенів свободи), тому F-критерій використовує відношення дисперсії, що пояснюється регресією (тобто величини SSR, поділеній на кількість незалежних змінних *k*), до дисперсії помилок [32 ,33,34, 35, 36, 37].

За визначенням F-статистика дорівнює середньому квадрату відхилень, обумовлених регресією (MSR), поділеній на дисперсію помилки (MSE):

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

де k - кількість незалежних змінних у регресійній моделі. Тестова статистика F має F-розподіл з *k* і *n-k-1* ступенями свободи.

При заданому рівні значущості α вирішальне правило сформулюється так: якщо , нульова гіпотеза відхиляється; в іншому випадку вона не відхиляється. Результати, оформлені у вигляді зведеної таблиці дисперсійного аналізу, наведені на рис. 30 [32 ,33,34, 35, 36, 37].

*Таблиця 3.1.*

Таблиця дисперсійного аналізу для перевірки гіпотези про статистичної значущості коефіцієнта регресії

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Джерело** | **Кількість ступенів свободи** | **Сума квадратів** | **Середньоквадратичне значення** | **F** |
| Регресія | k | SSR |  |  |
| Помилка | n-k-1 | SSE |  |  |
| Всього | n-1 | SST |  |  |

Використання t-критерію для коефіцієнта кореляції.

Раніше був введений коефіцієнт кореляції *r*, який представляє собою міру залежності між двома числовими змінними. З його допомогою можна встановити, чи існує між двома змінними статистично значуща зв'язок. Позначимо коефіцієнт кореляції між генеральними сукупностями обох змінних символом ρ. Нульова і альтернативна гіпотези формулюються наступним чином: (немає кореляції), (є кореляція). Перевірка існування кореляції [38 ,39,40, 41, 42, 43]:

|  |  |
| --- | --- |
| де |  |

Під час обговорення висновків, що стосуються нахилу генеральної сукупності, довірчі інтервали і критерії для перевірки гіпотез є взаємозамінними інструментами. Однак обчислення довірчого інтервалу, що містить коефіцієнт кореляції, виявляється більш складною справою, оскільки вид вибіркового розподілу статистики r залежить від істинного коефіцієнта кореляції [38 ,39,40, 41, 42, 43].

### 3.3.14. Оцінка математичного очікування і передбачення індивідуальних значень

У цьому розділі розглядаються методи оцінки математичного очікування відповіді *Y* і передбачення індивідуальних значень *Y* при заданих значеннях змінної *X [*38 ,39,40, 41, 42, 43].

Побудова довірчого інтервалу

У розділі про Метод найменших квадратів, регресійне рівняння дозволило передбачити значення змінної *Y* при заданому значенні змінної *X*. Отримана оцінка математичного очікування генеральної сукупності є точковою. Раніше для оцінки математичного очікування генеральної сукупності була запропонована концепція довірчого інтервалу. Аналогічно можна ввести поняття довірчого інтервалу для математичного очікування відповіді при заданому значенні змінної X [38 ,39,40, 41, 42, 43]:

|  |  |
| --- | --- |
|  |  |

де , - передбачене значення змінної *Y* при , - середньоквадратична помилка, *n* - обсяг вибірки, - задане значення змінної *X*, - математичне очікування змінної *Y* при ,

Аналіз формули (13) показує, що ширина довірчого інтервалу залежить від декількох факторів. При заданому рівні значимості зростання амплітуди коливань навколо лінії регресії, виміряний за допомогою середньоквадратичної помилки, призводить до збільшення ширини інтервалу. З іншого боку, як і слід було очікувати, збільшення обсягу вибірки супроводжується звуженням інтервалу. Крім того, ширина інтервалу змінюється в залежності від значень . Якщо значення змінної *Y* передбачається для величин *X* , близьких до середнього значення , довірчий інтервал виявляється краще, ніж при прогнозуванні відгуку для значень, далеких від середнього [38 ,39,40, 41, 42, 43].

Обчислення довірчого інтервалу для передбаченого значення.

Крім довірчого інтервалу для математичного очікування відповіді при заданому значенні змінної *X*, часто необхідно знати довірчий інтервал для передбаченого значення. Незважаючи на те що формула для обчислення такого довірчого інтервалу дуже схожа на формулу (13), цей інтервал містить передбачене значення, а не оцінку параметра. Інтервал для передбаченого відгуку YX = *Xi* при конкретному значенні змінної X i визначається за формулою [38 ,39,40, 41, 42, 43]:

|  |  |
| --- | --- |
|  |  |

де , - передбачене значення змінної *Y* при , - середньоквадратична помилка, *n* - обсяг вибірки, - задане значення змінної *X*, - математичне очікування змінної *Y* при ,

Підводні камені і етичні проблеми, пов'язані із застосуванням регресії

Труднощі, пов'язані з регресійним аналізом [38 ,39,40, 41, 42, 43]:

* Ігнорування умов застосовності методу найменших квадратів.
* Помилкова оцінка умов застосовності методу найменших квадратів.
* Неправильний вибір альтернативних методів при порушенні умов застосовності методу найменших квадратів.
* Застосування регресійного аналізу без глибоких знань про предмет дослідження.
* Екстраполяція регресії за межі діапазону зміни пояснює змінної.
* Плутанина між статистичної та причинно-наслідкового залежностями.

Широке поширення електронних таблиць і програмного забезпечення для статистичних розрахунків ліквідувало обчислювальні проблеми, що перешкоджали застосуванню регресійного аналізу. Однак це призвело до того, що регресійний аналіз стали застосовувати користувачі, що не володіють достатньою кваліфікацією і знаннями. Звідки користувачам знати про альтернативні методи, якщо багато хто з них взагалі не мають ні найменшого поняття про умови застосовності методу найменших квадратів і не вміють перевіряти їх виконання? Дослідник не повинен захоплюватися перемелюванням чисел - обчисленням зсуву, нахилу і коефіцієнта змішаної кореляції. Йому потрібні глибші знання [38 ,39,40, 41, 42, 43].

Отже, з точки зору регресійного аналізу всі ці набори даних абсолютно ідентичні. Якби аналіз був на цьому закінчено, ми втратили б багато корисної інформації. Про це свідчать діаграми розкиду (рис. 35) і графіки залишків (рис. 36), побудовані для цих наборів даних [38 ,39,40, 41, 42, 43].

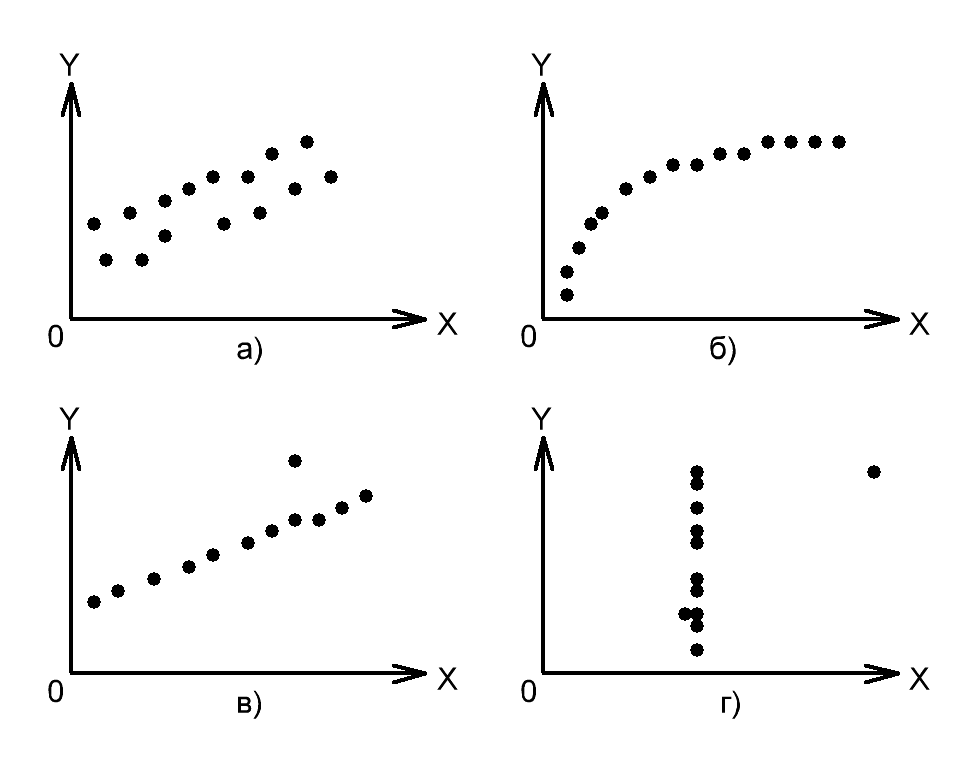


Рис. 3.8. Діаграми розкиду для чотирьох наборів даних

Діаграми розкиду і графіки залишків свідчать про те, що ці дані відрізняються одне від одного. Єдиний набір, розподілений вздовж прямої лінії, - рис а). Графік залишків, обчислених по рис. а), не має ніякої закономірності. Цього не можна сказати про рис б, в і г. Графік розкиду, побудований по рис б, демонструє яскраво виражену квадратичну модель. Цей висновок підтверджується графіком залишків, які мають параболічну форму. Діаграма розкиду і графік залишків показують, що набір даних *Y* містить викид. У цій ситуації необхідно виключити викид з набору даних і повторити аналіз. Метод, що дозволяє виявляти і виключати викиди зі спостережень, називається аналізом впливу. Після виключення викиду результат повторної оцінки моделі може виявитися зовсім іншим. Діаграма розкиду, побудована за даними з набору Г, ілюструє незвичайну ситуацію, в якій емпірична модель значно залежить від окремого відгуку. Такі регресійні моделі необхідно обчислювати особливо ретельно. Отже, графіки розкиду залишків є вкрай необхідним інструментом регресійного аналізу і повинні бути його невід'ємною частиною. Без них регресійний аналіз не заслуговує на довіру [38 ,39,40, 41, 42, 43].

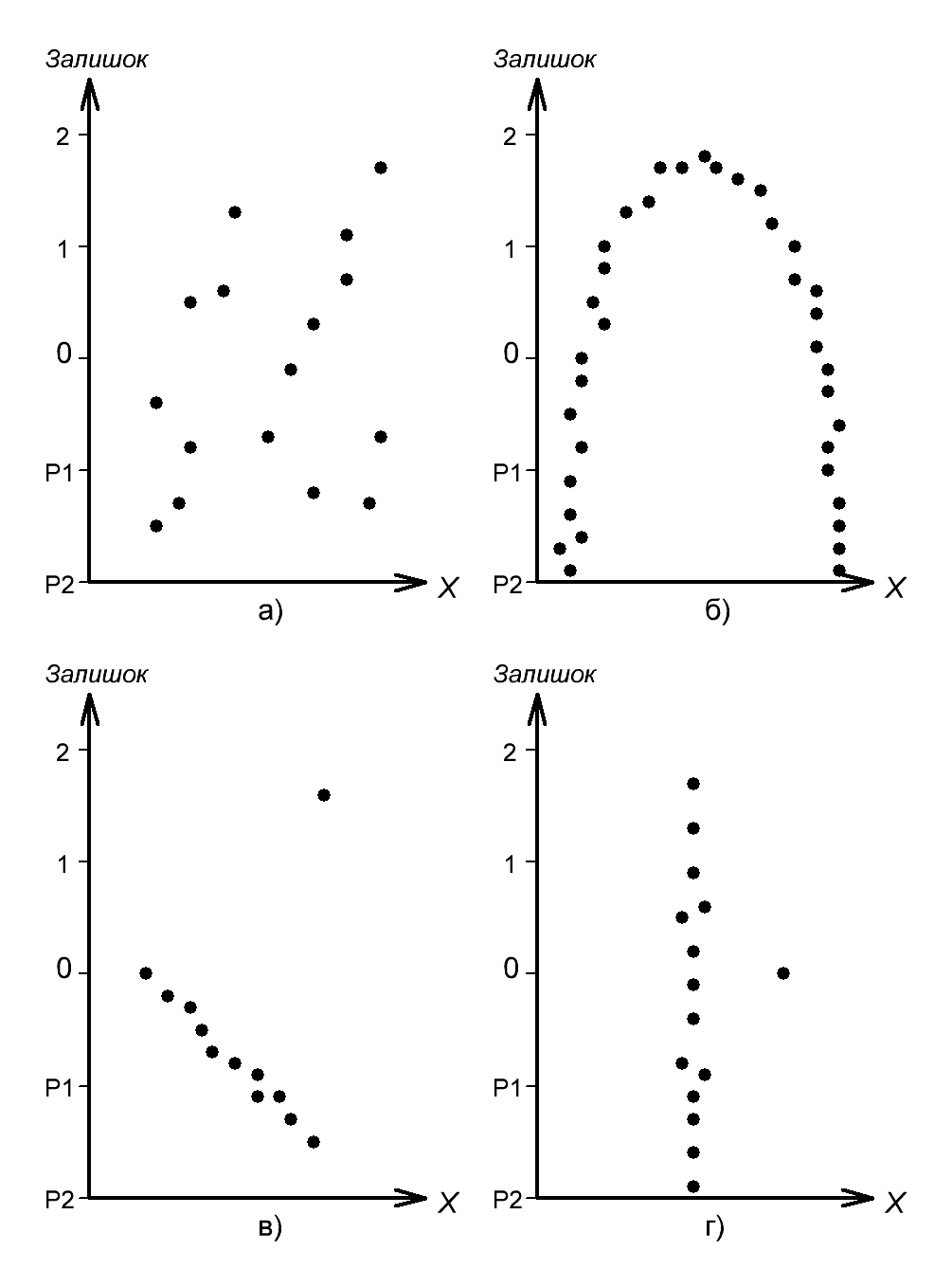


Рис. 3.9 Графіки залишків для чотирьох наборів даних

Як уникнути підводних каменів при регресійному аналізі [38 ,39,40, 41, 42, 43]:

* Аналіз можливого взаємозв'язку між змінними X і Y завжди починайте з побудови діаграми розкиду.
* Перш ніж інтерпретувати результати регресійного аналізу, перевіряйте умови його застосовності.
* Побудуйте графік залежності залишків від незалежної змінної. Це дозволить визначити, наскільки емпірична модель відповідає результатам спостереження, і виявити порушення сталості дисперсії.
* Для перевірки припущення про нормальний розподіл помилок використовуйте гістограми, діаграми «ствол і листя», блокові діаграми і графіки нормального розподілу.
* Якщо умови застосовності методу найменших квадратів не виконуються, використовуйте альтернативні методи (наприклад, моделі квадратичної або множинної регресії).
* Якщо умови застосовності методу найменших квадратів виконуються, необхідно перевірити гіпотезу про статистичної значущості коефіцієнтів регресії і побудувати довірчі інтервали, що містять математичне очікування і передбачене значення відгуку.
* Уникайте передбачати значення залежної змінної за межами діапазону зміни незалежної змінної.
* Майте на увазі, що статистичні залежності не завжди є причинно-наслідковими. Пам'ятайте, що кореляція між змінними не означає наявності причинно-наслідкового залежності між ними.

Як показано на структурній схемі (рис. 37), описано модель простої лінійної регресії, умови її застосування і способи перевірки цих умов. Розглянуто t -критерій для перевірки статистичної значущості нахилу регресії. Для передбачення значень залежної змінної використана регресійна модель. Розглянуто приклад, пов'язаний з вибором місця для торгової точки, в якому досліджується залежність річного обсягу продажів від площі магазину. Отримана інформація дозволяє точніше вибрати місце для магазину і передбачити його річний обсяг продажів. У наступних замітках буде продовжено обговорення регресійного аналізу, а також розглянуті моделі множинної регресії [38 ,39,40, 41, 42, 43].



Рис. 3.10. Структурна схема замітки

### 3.3.15. Критерій згоди «-квадрат»

Розподіл використовується для перевірки узгодженості набору даних з фіксованим розподілом ймовірностей. У критерії згоди частоти, що належать певній категорії, порівнюються з частотами, які є теоретично очікуваними, якби дані дійсно мали зазначений розподіл [44, 45, 46, 47, 48, 49].

Перевірка за допомогою критерію згоди виконується в кілька етапів. По-перше, визначається конкретний розподіл ймовірностей, яке порівнюється з вихідними даними. По-друге, висувається гіпотеза про параметри обраного розподілу ймовірностей (наприклад, про її математичне сподівання) або проводиться їх оцінка. По-третє, на основі теоретичного розподілу визначається теоретична ймовірність, відповідна кожної категорії. На закінчення, для перевірки узгодженості даних і розподілу застосовується тестова –Статистика [44, 45, 46, 47, 48, 49]:

|  |  |
| --- | --- |
|  |  |

де - спостережувана частота, - теоретична, або очікувана частота, *k* - кількість категорій, що залишилися після об'єднання, *р* - кількість оцінюваних параметрів.

Критерій для перевірки гіпотези від дисперсії або стандартного відхилення вважається класичною параметричної процедурою. При перевірці гіпотези про дисперсії генеральної сукупності або стандартного відхилення передбачається, що вихідні дані мають нормальний розподіл. Нажаль, -критерій досить чутливий до порушення цих припущень (тобто цей критерій не є стійким). Отже, якщо генеральна сукупність не має нормального розподілу, особливо, коли обсяг вибірки невеликий, точність критерію може значно знизитися [44, 45, 46, 47, 48, 49].

### 3.3.17. Класифікація на основі Байєсівського підходу

Байєсівський підхід до класифікації заснований на теоремі, яка стверджує, що якщо щільності розподілу кожного з класів відомі, то шуканий алгоритм можна виписати в явному аналітичному вигляді. Більш того, цей алгоритм оптимальний, тобто володіє мінімальною ймовірністю помилок [44, 45, 46, 47, 48, 49].

На практиці щільності розподілу класів, як правило, не відомі. Їх доводиться оцінювати (відновлювати) за навчальною вибіркою. В результаті байесовский алгоритм перестає бути оптимальним, так як відновити щільність по вибірці можна тільки з деякою погрішністю. Чим коротше вибірка, тим вище шанси підігнати розподіл під конкретні дані і зіткнутися з ефектом перенавчання [44, 45, 46, 47, 48, 49].

Байєсівський підхід до класифікації є одним з найстаріших, але до сих пір зберігає міцні позиції в теорії розпізнавання. Він лежить в основі багатьох досить вдалих алгоритмів класифікації [44, 45, 46, 47, 48, 49].

До числа байесовских методів класифікації відносяться:

* Наївний байесовский класифікатор
* Лінійний дискриминант Фішера
* Квадратичний дискриминант
* Метод парзеновского вікна
* Метод радіальних базисних функцій (RBF)
* Логістична регресія

### 3.3.17 Наївний байесовский класифікатор

Наївний байесівський класифікатор ( naїve Bayes ) - спеціальний окремий випадок байєсівського класифікатора , заснований на додатковому припущенні, що об'єкти описуються *n* статистично незалежними ознаками [50, 51, 52, 53, 54, 55]:

|  |  |
| --- | --- |
|  |  |

Припущення про незалежність означає, що функції правдоподібності класів представимо у вигляді [50, 51, 52, 53, 54, 55]:

|  |  |
| --- | --- |
|  |  |

де - щільність розподілу значень *j*-го ознаки для класу

Припущення про незалежність істотно спрощує завдання, тому що оцінити одновимірних щільності набагато легше, ніж одномірну щільність. На жаль, воно вкрай рідко виконується на практиці, звідси і назва методу [50, 51, 52, 53, 54, 55].

Наївний байесівський класифікатор може бути як параметричних, так і непараметричним, в залежності від того, яким методом відновлюються одномірні щільності [50, 51, 52, 53, 54, 55].

Основні переваги наївного байєсівського класифікатора - простота реалізації і низькі обчислювальні витрати при навчанні та класифікації. У тих рідкісних випадках, коли ознаки дійсно незалежні (або майже незалежні), наївний байесівський класифікатор (майже) оптимальний [50, 51, 52, 53, 54, 55].

Основний його недолік - відносно низька якість класифікації в більшості реальних задач [50, 51, 52, 53, 54, 55].

Найчастіше він використовується або як примітивний еталон для порівняння різних моделей алгоритмів, або як елементарний будівельний блок в алгоритмічних композиціях [50, 51, 52, 53, 54, 55].

### 3.3.19. Поняття кластеризації

Кластеризація (або кластерний аналіз) - це задача розбиття множини об'єктів на групи, які називаються кластерами. Усередині кожної групи повинні виявитися «схожі» об'єкти, а об'єкти різних групи повинні бути якомога більш відмінні. Головна відмінність кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий і визначається в процесі роботи алгоритму [56, 57, 58, 59, 60, 61].

Застосування кластерного аналізу в загальному вигляді зводиться до наступних етапів [56, 57, 58, 59, 60, 61]:

1. Відбір вибірки об'єктів для кластеризації.
2. Визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці. При необхідності - нормалізація значень змінних.
3. Обчислення значень міри схожості між об'єктами.
4. Застосування методу кластерного аналізу для створення груп схожих об'єктів (кластерів).
5. Представлення результатів аналізу.

Після отримання та аналізу результатів можливе корегування обраної метрики і методу кластеризації до отримання оптимального результату [56, 57, 58, 59, 60, 61].

Міри відстаней.

Отже, як же визначати «схожість» об'єктів? Для початку потрібно скласти вектор характеристик для кожного об'єкта - як правило, це набір числових значень, наприклад, зростання-вага людини. Однак існують також алгоритми, що працюють з якісними характеристиками [56, 57, 58, 59, 60, 61].

Після того, як ми визначили вектор характеристик, можна провести нормалізацію, щоб всі компоненти давали однаковий вклад при розрахунку «відстані». У процесі нормалізації все значення приводяться до деякого діапазону, наприклад, [-1, 1] або [0, 1] [56, 57, 58, 59, 60, 61].

Нарешті , для кожного пари об'єктів вимірюється «відстань» між ними - ступінь схожості. Існує безліч метрик, ось лише основні з них [56, 57, 58, 59, 60, 61]:

1. Евклідова відстань

Найбільш поширена функція відстані. Являє собою геометричних відстанню в багатовимірному просторі:

|  |  |
| --- | --- |
|  |  |

1. Квадрат евклідової відстані

Застосовується для додання більшої ваги більш віддаленим один від одного об'єктів. Це відстань обчислюється таким чином:

|  |  |
| --- | --- |
|  |  |

1. Відстань міських кварталів (Манхеттенська відстань)

Це відстань є середнім різниць по координатах. У більшості випадків ця міра відстані приводить до таких же результатів, як і для звичайного відстані Евкліда. Однак для цього заходу вплив окремих великих різниць (викидів) зменшується (тому що вони не зводяться в квадрат). Формула для розрахунку манхеттенської відстані [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

1. Відстань Чебишева

Це відстань може виявитися корисним, коли потрібно визначити два об'єкта як «різні», якщо вони розрізняються за якоюсь однією координаті. Відстань Чебишева обчислюється за формулою [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

1. Степенева відстань

Застосовується в разі, коли необхідно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Степенева відстань обчислюється за наступною формулою [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

де r і p - параметри, що визначаються користувачем. Параметр p відповідальний за поступове зважування різниць за окремими координатами, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри - r і p - дорівнюють двом, то це відстань збігається з відстанню Евкліда.

Вибір метрики повністю лежить на дослідника, оскільки результати кластеризації можуть істотно відрізнятися при використанні різних заходів [56, 57, 58, 59, 60, 61].

### 3.3.30. Класифікація алгоритмів кластеризації

Існують дві основні класифікації алгоритмів кластеризації.

1. Ієрархічні і плоскі.

Ієрархічні алгоритми (також звані алгоритмами таксономії) будують не одне розбиття вибірки на непересічні кластери, а систему вкладених розбиття. Тобто на виході ми отримуємо дерево кластерів, коренем якого є вся вибірка, а листям - найбільш дрібні кластера. Плоскі алгоритми будують одне розбиття об'єктів на кластери [56, 57, 58, 59, 60, 61].

1. Чіткі і нечіткі.

Чіткі (або непересічні) алгоритми кожному об'єкту вибірки ставлять у відповідність номер кластера, тобто кожен об'єкт належить тільки одного кластеру. Нечіткі (або пересічні) алгоритми кожному об'єкту ставлять у відповідність набір речових значень, що показують ступінь відносини об'єкта до кластерів. Тобто кожен об'єкт відноситься до кожного кластеру з певною ймовірністю [56, 57, 58, 59, 60, 61].

Об'єднання кластерів

У разі використання ієрархічних алгоритмів постає питання, як об'єднувати між собою кластера, як обчислювати «відстані» між ними [56, 57, 58, 59, 60, 61].

Існує кілька метрик:

1. Одиночна зв'язок (відстані найближчого сусіда).

У цьому методі відстань між двома кластерами визначається відстань між двома найбільш близькими об'єктами (найближчими сусідами) в різних кластерах. Результуючі кластери мають тенденцію об'єднуватися в ланцюжки [56, 57, 58, 59, 60, 61].

1. Повна зв'язок (відстань найбільш віддалених сусідів).

У цьому методі відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (тобто найбільш віддаленими сусідами). Цей метод зазвичай працює дуже добре, коли об'єкти походять з окремих груп. Якщо ж кластери мають подовжену форму або їх природний тип є «ланцюжковий», то цей метод непридатний [56, 57, 58, 59, 60, 61].

1. Незважене попарне середнє.

У цьому методі відстань між двома різними кластерами обчислюється що середня відстань між усіма парами об'єктів в них. Метод ефективний, коли об'єкти формують різні групи, проте він працює однаково добре і в випадках протяжних («ланцюжкового» типу) кластерів [56, 57, 58, 59, 60, 61].

1. Виважена попарне середнє.

Метод ідентичний методу невиваженого попарного середнього, за винятком того, що при обчисленнях розмір відповідних кластерів (тобто число об'єктів, що містяться в них) використовується в якості вагового коефіцієнта. Тому даний метод має бути використаний, коли передбачаються нерівні розміри кластерів[56, 57, 58, 59, 60, 61].

1. Метод незважений центроїда

У цьому методі відстань між двома кластерами визначається як відстань між їх центрами тяжкості.

1. Метод зваженого центроїда (медіана)

Цей метод ідентичний попередньому, крім того, що при обчисленнях використовуються ваги для обліку різниці між розмірами кластерів. Тому, якщо є або підозрюються значні відмінності в розмірах кластерів, цей метод виявляється переважно попереднього [56, 57, 58, 59, 60, 61].

Ієрархічні алгоритми, що використовують поняття порога

Альтернативою кластерним методам є ієрархічні алгоритми, що використовують поняття порога. Порогові алгоритми ефективні для вихідних сукупностей, у яких слабо виражений ланцюговий ефект і вони природно розпадаються на якусь кількість досить віддалених скупчень точок (кластерів).

Найбільш популярними серед таких алгоритмів є алгоритми типу FOREL ("формальний елемент"). FOREL є прикладом евристичного алгоритму класифікації, заснованого на ідеї об'єднання в один кластер об'єктів в областях їх найбільшого згущення. Таксони, одержувані цим алгоритмом, мають сферичну форму. Кількість таксонів залежить від радіуса сфер: чим менше радіус, тим більше виходить таксонів [56, 57, 58, 59, 60, 61].

### 3.3.31. Нейронні мережі Кохонена

Нейронні мережі Кохонена - типовий приклад нейромережевої архітектури, яка навчається без вчителя. Звідси і перелік вирішуваних ними завдань: кластеризація даних або прогнозування властивостей. Крім того, мережі Кохонена можуть використовуватися з метою зменшення розмірності даних з мінімальною втратою інформації [56, 57, 58, 59, 60, 61].

Раніше розглянуті архітектури нейронних мереж навчалися з учителем на вибірках даних, що включають безліч прикладів, що складаються з відповідних один одному пар вхідних і вихідних векторів. При цьому вихідні значення брали безпосередню участь в налаштуванні вагових коефіцієнтів. У нейронних мережах Кохонена вихідні вектора в навчальній вибірці можуть бути, але можуть бути і відсутніми, і, в будь-якому випадку, вони не беруть участі в процесі навчання. Тобто виходи не використовуються в якості орієнтирів при корекції синапсів. Саме тому даний принцип настройки нейронної мережі називається самонавчанням [56, 57, 58, 59, 60, 61].

У розглянутій архітектурі сигнал поширюється від входів до виходів в прямому напрямку. Структура нейронної мережі містить єдиний шар нейронів (шар Кохонена ) без коефіцієнтів зміщення (мал. 1). Загальна кількість вагових коефіцієнтів розраховується як добуток [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

Кількість нейронів дорівнює кількості кластерів, серед яких відбувається початкове розподіл і подальше перерозподіл навчальних прикладів. Кількість вхідних змінних нейронної мережі дорівнює числу ознак, що характеризують об'єкт дослідження і на основі яких відбувається віднесення його до одного з кластерів [56, 57, 58, 59, 60, 61].

Слід розрізняти власне самонавчання і самоорганізацію нейронної мережі Кохонена. При звичайному самонавчанні мережа має строго фіксовану структуру, тобто кількість нейронів, що не змінюються протягом усього життєвого циклу. При самоорганізованій мережі, навпаки, не має постійної структури. Залежно від знайденого відстані до нейрона-переможця, або цей нейрон використовується для кластеризації прикладу, або для поданого на входи прикладу створюється новий кластер з відповідними йому ваговими коефіцієнтами. Крім того, в процесі самоорганізації структури мережі Кохонена окремі нейрони можуть виключатися з неї [56, 57, 58, 59, 60, 61].

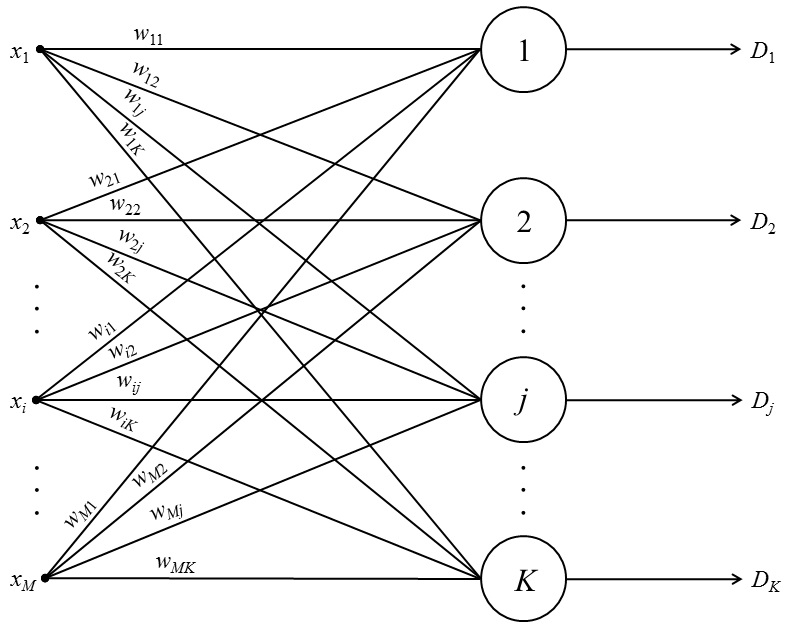


Рис. 3.11. Загальна структура нейронної мережі Кохонена

Нормалізація вхідних змінних виконується в межах [-1, 1] або [0, 1].

Для життєвого циклу нейронних мереж даної архітектури характерні три основні стадії життєвого циклу: навчання, кластерний аналіз та практичне використання [56, 57, 58, 59, 60, 61].

Алгоритм навчання мережі Кохонена включає етапи, склад яких залежить від типу структури: постійної (самонавчальна мережу) або змінною (самоорганізовану мережу). Для самонавчання послідовно виконуються:

1. Задання структури мережі (кількості нейронів шару Кохонена).
2. Випадкова ініціалізація вагових коефіцієнтів значеннями, що задовольняють одному з наступних обмежень:

* при нормалізації вихідної вибірки в межах [-1, 1]:

|  |  |
| --- | --- |
|  |  |

* при нормалізації вихідної вибірки в межах [0, 1]:

|  |  |
| --- | --- |
|  |  |

де *M* - кількість вхідних змінних мережі - характеристичних ознак об'єкта дослідження.

1. Подача на входи мережі випадкового навчального прикладу поточного навчання і розрахунок евклідових відстаней від вхідного вектора до центрів всіх кластерів:

|  |  |
| --- | --- |
|  |  |

1. За найменшому з значень вибирається нейрон-переможець *j*, в найбільшою мірою близький за значеннями з вхідним вектором. Для обраного нейрона (і тільки для нього) виконується корекція вагових коефіцієнтів:

|  |  |
| --- | --- |
|  |  |

де *v* - коефіцієнт швидкості навчання.

1. Цикл повторюється з кроку 3 до виконання одного або декількох умов закінчення:

* вичерпано заданий гранична кількість епох навчання;
* не відбулося значного зміни вагових коефіцієнтів в межах заданої точності протягом останньої доби навчання;
* вичерпано заданий граничний фізичний час навчання.

Коефіцієнт швидкості навчання може здаватися незмінним за межі [0, 1] або змінним значенням, поступово зменшується від циклу до циклу.

У разі самоорганізації мережі Кохонена алгоритм зазнає певних змін:

1. Здається критичну відстань , відповідне максимально допустимому евклидову відстані між входами прикладу і вагами нейрона-переможця. Початкова структура не містить нейронів. При подачі на входи мережі самого першого прикладу навчальної вибiрки створюється перший нейрон з ваговими коефіцієнтами, рівними поданням вхідним значенням [56, 57, 58, 59, 60, 61].
2. На входи мережі подається новий випадково обраний приклад поточної епохи навчання, розраховуються евклідові відстані від прикладу до центру кожного кластера по співвідношенню (3) і визначається нейрон-переможець з найменшим з них .
3. Якщо виконується умова , проводиться корекція вагових коефіцієнтів відповідного нейрона-переможця по співвідношенню (4), в іншому випадку в структуру мережі додається новий нейрон, вагові коефіцієнти якого приймаються чисельно рівними вхідним значенням поданого прикладу.
4. Процедура повторюється з п.3. Якщо протягом останньої доби навчання будь-які кластери залишилися не задіяними, відповідні нейрони виключаються зі структури мережі Кохонена [56, 57, 58, 59, 60, 61].
5. Обчислення закінчуються, якщо виконується одна з умов, прописані в алгоритмі самонавчання мережі фіксованого структури [56, 57, 58, 59, 60, 61].

Ще одна модифікація алгоритмів самонавчання і самоорганізації передбачає корекцію вагових коефіцієнтів не тільки нейрона-переможці, а й усіх інших нейронів. Для цього слід використовувати коефіцієнт швидкості навчання, регресний зі збільшенням відстані до центру кластера *Rj* :

|  |  |
| --- | --- |
|  |  |

де - критичне значення відстані: чим воно менше, тим більш значні будуть коректування ваг найближчих до навчального наприклад кластерів і практично незначущі - ваг більш-менш віддалених від нього; *β* - параметр, який встановлює ступінь нелінійності впливу відстані на коефіцієнт швидкості; - базове (максимально можливе) значення коефіцієнта швидкості на поточній епосі навчання [56, 57, 58, 59, 60, 61].

Як значення можна розраховувати середнє відстань для кожного кластера при поточному пред'явленні навчального прикладу. Параметр *β* рекомендується вибирати рівним 3,0 ± 0,5.

Як правило, практично під час використання самоорганізації нейронної мережі Кохонена доводиться стикатися ще з однією проблемою. З одного боку, якісь кластери можуть містити занадто маленька кількість прикладів, що призводить до складнощів у наступному узагальненні інформації. З іншого боку, деякі кластери можуть виявитися занадто великими, тобто містити дуже багато прикладів. В цьому випадку для регулювання розміру кластера і вирішення проблеми його переповненості можна задати в якості додаткового параметра граничне число прикладів, які формують кластер *Nпр*. Якщо в якийсь момент виявляється, що новий приклад повинен бути віднесений до кластеру, розмір якого вже максимальний, приймається рішення про створення іншого кластера, центр якого буде представляти собою вектор змінних одного з *Nпр+1* прикладів кластера (включаючи новий) найбільш віддаленого від центру даного кластера [56, 57, 58, 59, 60, 61].

До навченої нейронної мережі застосовується процедура кластерного аналізу - процедури опису властивостей кластера на основі аналізу кількісного і якісного складів прикладів, які сформували його. Слід враховувати, що опис кластерів може базуватися не тільки на значеннях вхідних змінних навчальної вибірки, а й на значеннях змінних, які не брали участі у формуванні кластерів. Зокрема, в опис можуть входити дані про середні значення таких змінних серед всіх прикладів, які сформували кластер. Крім того, доцільно для кожного кластера мати дані про середньоквадратичному відхиленні або дисперсії по кожній змінній [56, 57, 58, 59, 60, 61].

При практичному використанні нейронної мережі Кохонена новий приклад подається на її вхід і відноситься до одного з існуючих кластерів, або робиться висновок про неможливість такого віднесення (при великій відстані до центру найближчого кластера). Якщо вибір кластера відбувся, його опис, отримане в результаті кластерного аналізу, і відповідні кластеру рішення повинні поширюватися в тому числі на поданий приклад [56, 57, 58, 59, 60, 61].

Практичне використання мережі Кохонена полегшується за рахунок візуалізації результатів кластеризації. В результаті самонавчання (самоорганізації) мережі виходить набір кластерів, кожен з яких характеризується своїм центром (значеннями вагових коефіцієнтів відповідного нейрона) і кількістю навчальних прикладів, які сформували його. Чи не складає ніяких труднощів визначити евклідова відстань між центрами всіх можливих пар кластерів і графічно зобразити їх на так званій карті Кохонена - двовимірної графічної структурі, що дозволяє судити не тільки про розміри і положення кожного окремо взятого кластера, а й про близькість один до одного і взаємне розташування окремих кластерів [56, 57, 58, 59, 60, 61].

### 3.3.33. Кластеризація методом k-means

Метод *k-*means - це спеціальний алгоритм кластеризації, що має на увазі, що у нас є масив даних, які ми хочемо згрупувати в кластери, а точніше - в *k*кластерів [56, 57, 58, 59, 60, 61].

Вхідними даними в методі *k-*means є тільки матриця наших *X*. Як правило, ми формуємо її так, щоб кожен рядок представляла окремий приклад (зразок), а кожен стовпець - окремий ознака або, користуючись термінами з статистики, фактор. Зазвичай ми говоримо, що є *N*прикладів і *D*ознак, так що *X*є матрицею розмірності *N*x*D* [56, 57, 58, 59, 60, 61].

В алгоритмі методу *k-*means є два основних етапи. Спочатку ми вибираємо *k*різних центрів кластерів - як правило, це просто випадкові точки в наборі даних. Потім ми переходимо до нашого основного циклу, який також складається з двох етапів. Перший - це вибір, до якого з кластерів належить кожна точка з *X*. Для цього ми беремо кожен приклад і вибираємо кластер, чий центр ближче всього. Не забувайте, що спочатку ми вибираємо центри випадковим чином. Другий етап - заново обчислити кожен центр кластера, ґрунтуючись на безлічі точок, які до нього приписані. Для цього беруться всі відповідні приклади і обчислюється їх середнє значення, звідси і назва методу - «метод *k-*means». Все це робиться до тих пір, поки алгоритм не зійде, тобто поки не припиниться зміна в розподілі точок по кластерам або в координатах центрів кластерів. Як правило, це відбувається дуже швидко - в районі від 5 до 14 проходів циклу. Це сильно відрізняється від градієнтного спуску в глибокому навчанні, де можуть пройти тисячі ітерацій, поки не відбудеться сходження [56, 57, 58, 59, 60, 61].

Розглянемо наочний приклад роботи методу *k*-середній. На першому етапі ми приписуємо центри кластерів *m1*і *m3*до випадкових точках *X*. На наступному етапі - що є першим етапом основного циклу - ми вибираємо, до якого з кластерів належить кожна точка. Так, на малюнку дві точки зліва належать до кластеру 1, а дві точки праворуч - до кластеру 3. На наступному етапі перераховуємо середні значення *m1*і *m3*. Тоді *m1*виходить середнім значенням двох точок зліва, оскільки вони належать цьому кластеру. Все, алгоритм зійшовся, оскільки більше не може бути змін до присвоєння даних до кластерів і, отже, в значеннях *m1*і *m3* [56, 57, 58, 59, 60, 61].

М'який метод *k-*means

При реалізації методу *k-*means ми можемо виявити, що він дуже чутливий до ініціації. Нагадаю, що для ініціації середніх значень ми просто вибирали ряд випадкових точок в наборі даних. Один з часто використовуваних варіантів - запустити метод *k*-середній кілька разів і використовувати той результат, який дає нам найкраще кінцеве значення функції витрат - як її обчислювати, ми обговоримо на наступній лекції [56, 57, 58, 59, 60, 61].

Це говорить, що функція витрат чутлива до локальних мінімумів. Один із способів подолати цю перешкоду - використання «нечіткого» членства в кожному класі. Це означає, що кожна точка наших даних не відноситься повністю до того чи іншого класу, а становить суму приладдя до класів, наприклад, може належати на 60% кластеру 1 і на 40% кластеру 3. Можна побачити, що можна отримати м'який метод *k-*means, лише трохи підправив алгоритм «звичайного» методу *k-*means [56, 57, 58, 59, 60, 61].

Отже, перша частина в «м'якому» варіанті залишається незмінною - ми ініціюємо *k*центрів кластерів в випадкових точках набору даних. Зміни починаються в тілі головного циклу. Так, першим етапом стає обчислення меж розподілу кластерів [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

Як ви можете переконатися, тепер наше завжди буде дробом, числом між 0 і 1, тоді як в «жорсткому», або «звичайному», методі *k-*means завжди дорівнює точно нулю або одиниці [56, 57, 58, 59, 60, 61].

Другий етап дуже схожий - ми лише перераховуємо середні значення, ґрунтуючись на кордонах розподілах кластерів:

|  |  |
| --- | --- |
|  |  |

Як бачите, виходить щось схоже на обчислення середньозваженого значення - більш високе значення означає, що дане в більшій мірі відноситься до кластеру *k*, що, в свою чергу, означає, що воно має більший вплив на обчислення відповідного середнього значення [56, 57, 58, 59, 60, 61].

Цільова функція методу *k-*means

Як і в разі навчання з учителем, вельми важливо визначити цільову функцію, яку ми намагаємося мінімізувати. Припускаючи, що ми використовуємо в якості міри відстані евклідова відстань, цільова функція буде мати вигляд [56, 57, 58, 59, 60, 61]:

|  |  |
| --- | --- |
|  |  |

Таким чином, в дійсності це лише квадрат відстані, зважений на величину приналежності *x*, тобто якщо *x(n)*знаходиться далеко від центру кластера *k*, то можна припускати, що його вплив на кластер буде дуже малим [56, 57, 58, 59, 60, 61].

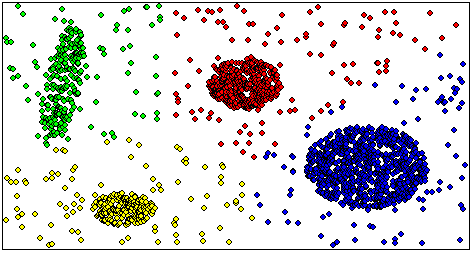


Рис. 3.13. Результат класифікації методами *k-*means

### 3.3.34. EМ - масштабований алгоритм кластеризації

Кластеризація є однієї з найбільш важливих завдань Data Mining. В даний час розроблено велику кількість методів і алгоритмів кластеризації але, на жаль, не всі вони можуть ефективно працювати з великими масивами даних, тому подальші дослідження в цьому напрямку пов'язані з подоланням цієї проблеми. Одним з широко відомих в аналітичному співтоваристві алгоритмів кластеризації, що дозволяють ефективно працювати з великими обсягами даних, є EM-алгоритм. Його назва походить від слів "expectation-maximization", що перекладається як "очікування-максимізація". Це пов'язано з тим, що кожна ітерація містить два кроки - обчислення математичних очікувань ( expectation ) і максимізацію ( maximisation ). Алгоритм заснований на методиці ітеративного обчислення оцінок максимальної правдоподібності, запропонованої в 1977 р (AP Demster , NM Laird , DB Rubin . Maximum Likelihood from Incomplete Data via the EM Algorithm ) [62].

В основі ідеї EM-алгоритму лежить припущення, що досліджуване безліч даних може бути змодельоване за допомогою лінійної комбінації багатовимірних нормальних розподілів, а метою є оцінка параметрів розподілу, які максимізують логарифмічну функцію правдоподібності, використовувану в якості запобіжного якості моделі. Іншими словами, передбачається, що дані в кожному кластері підкоряються певним законом розподілу, а саме, нормальному розподілу (рис). З урахуванням цього припущення можна визначити параметри - математичне очікування і дисперсію, які відповідають закону розподілу елементів в кластері, найкращим чином "невластивому" до спостережуваних даними [62].



Рис. 3.14. Розподіл елементів в кластерах

Таким чином, ми припускаємо, що будь-який спостереження належить до всіх кластерів, але з різною ймовірністю. Тоді завдання полягатиме в "підгонці" розподілів суміші до даних, а потім у визначенні ймовірностей приналежності спостереження до кожного кластеру. Очевидно, що спостереження повинно бути віднесено до того кластеру, для якого ця можливість вище [62].

Серед переваг EM-алгоритму можна виділити наступні [62]:

* Потужна статистична основа.
* Лінійне збільшення складності при зростанні обсягу даних.
* Стійкість до шумів та перепустками в даних.
* Можливість побудови бажаного числа кластерів.
* Швидка збіжність при вдалій ініціалізації [62].

Однак алгоритм має і ряд недоліків. По-перше, припущення про нормальність всіх вимірювань даних не завжди виконується. По-друге, при невдалій ініціалізації збіжність алгоритму може виявитися досить повільним. Крім цього, алгоритм може зупинитися в локальному мінімумі і дати квазіоптимальне рішення [62].

Статистичні основи алгоритму

Як зазначалося вище, EM-алгоритм передбачає, що кластеризуємі дані підкоряються лінійній комбінації (суміші) нормальних розподілів. Щільність ймовірності нормального розподілу має вигляд [62]:

|  |  |
| --- | --- |
|  |  |

де *μ* - математичне очікування, дисперсія.

Багатовимірне нормальний розподіл для *q*-мірного простору є узагальненням попереднього виразу. Багатовимірна нормальна щільність для *q*-мірного вектора може бути записана у вигляді [62]:

|  |  |
| --- | --- |
|  |  |

де *Σ* - коваріаційна матриця розміром *q×q*, яка, як відомо, є узагальненням дисперсії для багатовимірної випадкової величини, *μ* вдає із себе *q*-мірний вектор математичних очікувань, - визначник ковариационной матриці, *T*- оператор транспонування.

|  |  |
| --- | --- |
|  |  |

Алгоритм передбачає, що дані підкоряються суміші багатовимірних нормальних розподілів для *q* змінних. Модель, що представляє собою суміш гаусових розподілів задається у вигляді [62]:

|  |  |
| --- | --- |
|  |  |

де - нормальний розподіл для *i*-го кластера, - частка (вага) ii-го кластера в вихідній базі даних.

Існують два підходи до вирішення завдань кластеризації: заснований на відстані і заснований на щільності. Перший підхід полягає у визначенні областей простору ознак, усередині яких точки даних розташовані ближче один до одного, ніж до точок інших областей, щодо деякої функції відстані (наприклад, евклідової). Другий - виявляє області, які є більше "заселеними", ніж інші. Алгоритми кластеризації можуть працювати зверху вниз (ієрархічно) і від низу до верху (агломеративно). Агломеративні алгоритми, як правило, є точнішими, хоча і працюють повільніше [62].

Алгоритм EM заснований на обчисленні відстаней. Він може розглядатися як узагальнення кластеризації на основі аналізу суміші імовірнісних розподілів. У процесі роботи алгоритму відбувається ітеративне поліпшення рішення, а зупинка здійснюється в момент, коли досягається необхідний рівень точності моделі. Мірою в даному випадку є монотонно збільшується статистична величина, звана логарифмічною правдоподібністю. Метою алгоритму є оцінка середніх значень *C*, коваріацій *R* і ваг суміші *W* для функції розподілу ймовірності, описаної вище. Параметри, оцінені алгоритмом, зберігаються в таблиці виду [62]:

*Таблиця 3.3.*

**Параметри, оцінені алгоритмом ЕМ**

|  |  |  |
| --- | --- | --- |
| Матриця | Розмір | Містить |
| C | q×k | Математичні очікування, *μ* |
| R | q×q | Коваріації, *Σ* |
| W | k×1 | Ваги, |

Слід зазначити, що одна з популярних алгоритмів кластеризації k-means є окремим випадком алгоритму EM, коли *W* і *R* постійні.

Алгоритм починає роботу з ініціалізації, тобто деякого наближеного рішення, яке може бути вибрано випадково або задано користувачем виходячи з деяких апріорних відомостей про вихідних даних. Найбільш загальним способом ініціалізації є привласнення елементам матриці математичних очікувань випадкових значень *C←μ* Random, початкова коваріаційна матриця визначається як одинична *r←I*, ваги кластерів задаються однаковими [62].

Слід звернути увагу, що алгоритм може "застрягти" в локальному оптимумі і дати квазіоптимальне рішення при виборі невдалого початкового наближення. Тому одним з його недоліків слід вважати чутливість до вибору початкового стану моделі [62].

Для оптимізації використовуваного обсягу пам'яті, алгоритм може працювати в двох режимах. У першому завантажується тільки частина доступних даних і на їх основі робиться спроба побудови моделі. Якщо вона увінчалася успіхом, то алгоритм завершує роботу, в іншому випадку завантажується наступна порція даних і т.д., поки не будуть отримані прийнятні результати. У другому режимі завантажуються відразу всі наявні дані. Як правило, останній варіант забезпечує більш точну підгонку моделі, але пред'являє більш жорсткі вимоги до обсягу оперативної пам'яті [62].

Чисельний експеримент

Для ілюстрації роботи алгоритму EM і його порівняння з k-means розглянемо результати чисельного експерименту, для проведення якої була взята вибірка, представлена на малюнку 3 [62].

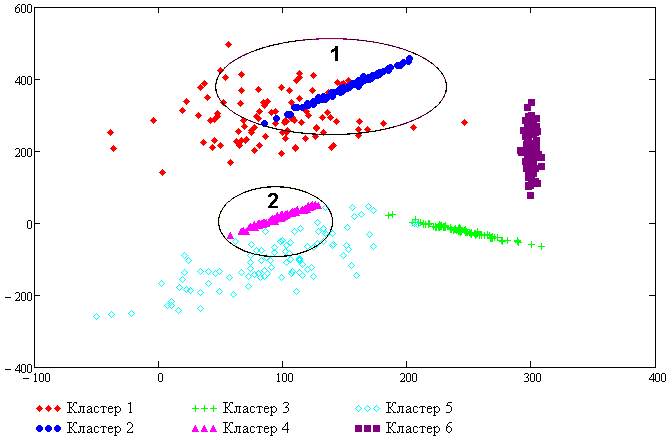


Рис. 3.14. Вихідні кластери

Зверніть увагу, що вихідний набір даних не є простим з точки зору завдання кластеризації, оскільки є явне перекриття кластерів (області 1 і 3). В області 1 перекриваються кластери 1 і 3, а в області 3 кластери 4 і 5. Кластери 3 і 6 розташовані відокремлено і, як очікується, будуть легко розділено [62].

Для алгоритму k - means особливі труднощі повинні виникнути в місцях перекриття кластерів. Дане припущення підтверджується результатами, представленими на рис. [62].

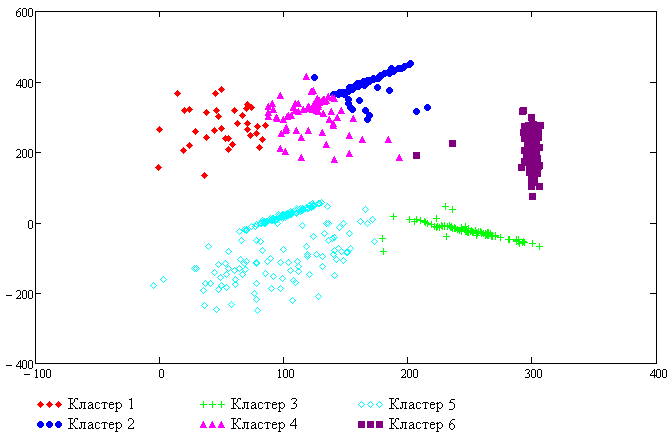


Рис. 3.15. Результати кластеризації k- means

У місцях перекриття кластерів спостерігається найбільше число помилок. У той же час відокремлені кластери 3 і 6 були розпізнані алгоритмом k - means без помилок. Як можна побачити на малюнку 4, алгоритм EM вельми успішно виявив перекриваються кластери, хоча й майже не розпізнав кластер 6 [62].

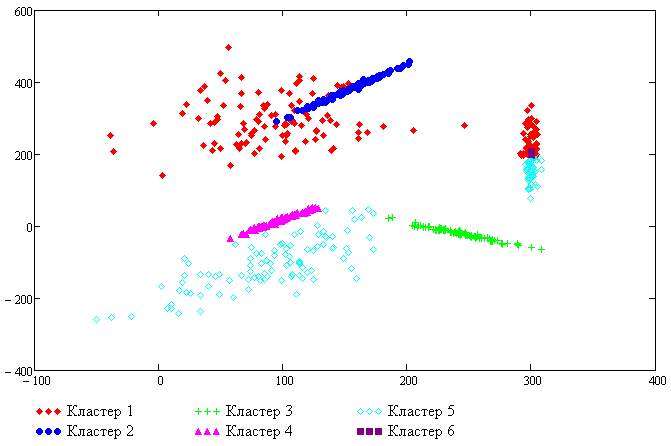


Рис. 3.17. Результати кластеризації EM

Таким чином, можна зробити висновок, що алгоритм k - means може мати перевагу при роботі з відокремленими (неперекривающіхся) кластерами, але повністю програє алгоритму EM при наявності їх перекриття [62].

## 3.4. Висновок до розділу

В боротьбі за безпеку інформаційної системи ефективність мають різноманітні методи аналізу кіберзагроз. Традиційні методи аналізу такі як пряма експертна оцінка, статистичний аналіз та факторний аналіз мають недостатню ефективність при постійному аналізі великих об’ємів даних. На допомогу традиційним методам аналізу приходять методи машинного навчання та розпізнавання образів, які дають можливість виявити нові фактори аналізу кіберзагроз для їх ефективної та достовірної оцінки.

Отже для ефективного захисту інформаційного простору необхідно удосконалити методологію аналізу кіберзагроз методами машинного навчання у поєднанні з статистичними методами відкритих каналів зв’язку на предмет прихованої інформації, яка може завдати шкоду інформаційному простору.

Розглянуто модель простої лінійної регресії, умови її застосування і способи перевірки цих умов. Розглянуто t -критерій для перевірки статистичної значущості нахилу регресії. Для передбачення значень залежної змінної використана регресійна модель. Розглянуто приклад, пов'язаний з вибором місця для торгової точки, в якому досліджується залежність річного обсягу продажів від площі магазину. Отримана інформація дозволяє точніше вибрати місце для магазину і передбачити його річний обсяг продажів. У наступних замітках буде продовжено обговорення регресійного аналізу, а також розглянуті моделі множинної регресії. Також розглянуто критерій розподілу «Хі-квадрат» Пірсона, а також розглянуто методи кластерного аналізу. Показано переваги та недоліки даних алгоритмів.

Також розглянуто методи приховування кіберзагроз та стеганоаналізу, що дозволяє показати принципи їх застосування в практичних цілях в даній роботі. Так як все більше і більше розробників шкідливого ПЗ починає використовувати стеганографію, в тому числі - для приховування комунікації з командним центром і для завантаження модулів. Це дає результат, адже процедури аналізу контейнерів імовірнісні і дорогі, отже, більшість захисних рішень не можуть собі дозволити обробляти всі об'єкти, які потенційно можуть бути заповненими контейнерами.

# Розділ 3. Алгоритм оцінки наявності прихованої інформації в каналі зв’язку

## 3.1. Причини використання стеганографії

Одним з найбільш перспективних підходів для виявлення факту існування прихованого каналу передачі інформації є підхід, який представляє введення в файл приховується як порушення статистичних закономірностей природних контейнерів. При цьому підході аналізуються статистичні характеристики досліджуваної послідовності і встановлюється, чи схожі вони на характеристики природних контейнерів (якщо так, то прихованої передачі інформації немає), або вони схожі на характеристики стегоконтейнера (якщо так, то виявлено факт існування прихованого каналу передачі інформації). Цей клас стегоатак є імовірнісним, тобто вони не дають однозначної відповіді, а формують оцінки типу «дана досліджувана послідовність з ймовірністю 90% містить приховане повідомлення». Імовірнісний характер статистичних методів стегоаналізу не є істотним недоліком, так як на практиці ці методи часто видають оцінки ймовірності існування стегоканалу, що відрізняються від одиниці або нуля на нескінченно малі величини [63,64].

За недавній час ми спостерігали використання стеганографії в наступних шкідливі програми і засобах кібершпіонажу:

* Microcin (AKA six little monkeys);
* NetTraveler;
* Zberp;
* Enfal (its new loader called Zero.T);
* Shamoon;
* KinS;
* ZeusVM;
* Triton (Fibbit).

Три головні причини використання стеганографії [63,64]:

* Це дозволяє їм приховати сам факт завантаження / вивантаження даних, а не тільки самі дані [63,64].;
* Допомагає обійти DPI-системи, що актуально в корпоративних мережах [63,64].;
* Використання стеганографії може дозволити обійти перевірку в AntiAPT-продуктах, оскільки останні не можуть обробляти всі графічні файли (їх занадто багато в корпоративних мережах, а алгоритми аналізу досить дорогі).
* Проблеми в наявності [63,64].:
* Використання стеганографії сьогодні - дуже популярна ідея серед авторів шкідливого і шпигунського ПЗ [63,64].;
* Антивірусні засоби взагалі і засоби захисту периметра зокрема мало що можуть зробити з заповненими контейнерами: їх дуже важко виявити оскільки вони виглядають як звичайні графічні (і не тільки) файли [63,64].;
* Всі існуючі програми для детектування стеганографії по суті є PoC- Proof-of-Concept, і їх логіка не може бути імплементована в промислові засоби захисту через низьку швидкості роботи, не надто високого рівня детектування, і іноді навіть іноді - через помилок в математиці (ми бачили і такі випадки) [63,64]..

Для непомітного вбудовування даних стеганокодер повинен вирішити три задачі: виділити підмножину біт, модифікація яких мало впливає на якість (незначущі біти), вибрати з цієї підмножини потрібну кількість біт відповідно до розміру прихованого повідомлення і виконати їх зміну. Якщо статистичні характеристики контейнеру не змінилися, то вбудовування інформації можна рахувати успішним. Так як розподіл незначущих біт часто близько до білого шуму, то вбудовані дані повинні мати той же характер. Це досягається за рахунок попереднього шифрування повідомлення або його стиснення [13].

Стеганоаналітік на основі вивчення сигналу завжди може виділити підмножину незначущих біт, роблячи ті ж припущення, що і в стеганографії. Далі він повинен перевірити відповідність їх статистики передбачуваної. При цьому, якщо аналітик має у своєму розпорядженні кращу модель, даних, ніж стеганограф, то вкладення буде виявлено. Тому, по-справжньому хороші моделі сигналів різного характеру, ймовірно, тримаються в секреті, і ви не зустрінете їх у відкритих публікаціях. Можна лише дати рекомендації загального характеру. При побудові моделі треба враховувати [13].:

* неоднорідність послідовностей відліків;
* залежність між бітами в відліки (кореляцію);
* залежність між відліками;
* статистику довжин серій (послідовностей з однакових біт).

Відповідність реально спостерігається статистики до очікуваної, зазвичай перевіряється за допомогою критерію хі-квадрат. Перевірка може здійснюватися на рівні монобітів, дібітів і т.д. Можливі й більш складні тести, аналогічно застосовується при тестуванні криптографічно безпечних генераторів випадкових чисел. Як показано в одній з робіт на прикладі звукових файлів, критерій хі-квадрат дозволяє виявити модифікацію всього лише 10% незначущих бітів [13]..

У порівнянні з досить добре дослідженими криптографічними системами, поняття і оцінки безпеки стеганографічних систем більш складні і допускають більше число їх тлумачень. Це пояснюється як недостатньою теоретичної та практичної опрацюванням питань безпеки стеганосистем, так і великою різноманітністю завдань стеганографічного захисту інформації [13]..

Тому ціллю даної роботи є створення системи, яка може аналізувати потік даних в каналі зв’язку, не розшифровувати, чи виділяти її з каналу передачі, а лише виявити факт передачі інформації. Під каналом зв’язку в цій роботі буде тлумачитися статичне зображення (одне або декілька). Для визначення факту передачі будемо застосовувати методи машинного навчання а також теорію розпізнавання образів [13].

## 3.3. Статистична оцінка двох класів контейнерів

Для дослідження скористаємося двома типами контейнерів, їх гістограмами і бітовими зрізами найменш значимого біта [65].

|  |  |
| --- | --- |
| 4 | C:\Users\talav\AppData\Local\Microsoft\Windows\INetCache\Content.Word\4.bmp |
| Рис. 3.1. Порожній контейнер | Рис. 3.3. Заповнений контейнер |
| C:\Users\talav\AppData\Local\Microsoft\Windows\INetCache\Content.Word\GrayNoStego.bmp | E:\Diplom\Image\GrayStego.bmp |
| Рис. 3.3 Півтонове уявлення порожнього контейнера | Рис. 3.4. Півтонове уявлення заповненого контейнера |
| E:\Diplom\Image\Test480_4.bmp | E:\Diplom\Image\TestStego480_4.bmp |
| Рис. 3.5. Зріз найменш значимого біта червоного спектра порожнього контейнера | Рис. 3.6. Зріз найменш значимого біта червоного спектра заповненого контейнера |

Контейнер в загальному випадку можна визначити як двомірну функцію *I(i, j)* де *i* і *j* - це просторові (площинні) координати, а амплітуда *I* для кожної пари координат *(i, j)* називається інтенсивністю або яскравістю контейнера в точці з цими координатами. Самі координати *(i, j)* називаються пікселями [65].

Для імовірнісних методів аналізу даних необхідно від гістограми перейти до розподілу випадкової величини (Рис. ). Гістограма *H(x)* (1) групує відліки, що мають однакові величини, разом. Це дозволяє обчислювати статистику, працюючи з групами, а не з великим числом окремих відліків [65].

|  |  |
| --- | --- |
| *G*- функція угруповання вибірки випадкових величин; | (1) |

Таким чином, верхня частина контуру гістограми утворює статистичний аналог для щільності ймовірності, так само, як і емпірична функція є статистичним аналогом для функції розподілу [65].

Інтеграл від щільності ймовірності Ф(х) (3) використовується для знаходження ймовірності знаходження сигналу всередині певної області величин. Це робить інтеграл від щільності розподілу ймовірності досить важливим для класифікації, тому необхідно його розглядати в цьому дослідженні [65].

|  |  |
| --- | --- |
|  | (3) |

Як видно з дослідження гістограма приймає форму нормального закону розподілу, де інтегральна функція розподілу яркостей пікселів контейнера має вигляд (3) і зображена на рис. 9.

|  |  |
| --- | --- |
|  | (3) |

При детальному розгляді графіків видно, що вони трохи зміщені відносно один одного, це свідчить про те, що змінюються статистичні характеристики контейнера при встановленні шкідливої ​​інформації. Це легко проконтролювати оцінками математичного очікування, дисперсії і середньоквадратичного відхилення [65].

Таблиця 3.1.

**Статистичні характеристики контейнерів**

|  |  |  |
| --- | --- | --- |
|  | Порожній контейнер | Заповнений контейнер |
| Математичне очікування | 0.5150 | 0.5350 |
| дисперсія | 0.0317 | 0.0337 |
| середньоквадратичне відхилення | 0.1194 | 0.1160 |

З гістограм яскравостей пікселів, законів розподілу та статистичних характеристик, які отримані в ході експериментів, можна зробити висновок в їх відмінностях, на основі яких і буде проводитися подальша оцінка і класифікація контейнерів.

Рис. 3.7. Гістограма порожнього і заповненого контейнера з розподілом коефіцієнтів яскравості пікселів.

Рис. 3.8. Інтегральна оцінка розподілу ймовірності коефіцієнтів яскравості пікселів Ф (х) заповненого (помаранчевий) і порожнього (синій) контейнерів, де x - координати коефіцієнтів яскравості пікселя.

Проведемо заповнення порожнього контейнера корисною інформацією, з різним відсотком заповнення. В одному випадку глибина зміни бітів буде тільки 1 біт, а в іншому випадку – 3 біти. Обчислимо статистичні характеристики: математичне очікування, дисперсія, середньоквадратичне відхилення та побудуємо графіки зміни цих характеристик відповідно до відсотка заповнення контейнера.

*Таблиця 3.3.*

**Результати оцінки контейнерів**

| Відсоток заповнення | Глибина заповнення - 1 біт | | Глибина заповнення - 3 біт | |
| --- | --- | --- | --- | --- |
| Математичне очікування | | Середньоквадратичне відхилення | |
| 1 | 3 | 3 | 4 | 5 |
| 1 | 0,385456 | 0,38545 | 0,343058 | 0,343063 |
| 3 | 0,385455 | 0,385444 | 0,343059 | 0,343013 |
| 3 | 0,385453 | 0,385437 | 0,343059 | 0,343017 |
| 4 | 0,385451 | 0,385433 | 0,343061 | 0,343014 |
| 5 | 0,385449 | 0,385435 | 0,343063 | 0,343014 |
| 6 | 0,385448 | 0,385415 | 0,343063 | 0,343089 |
| 7 | 0,385446 | 0,385413 | 0,343065 | 0,343093 |
| 8 | 0,385445 | 0,385404 | 0,343013 | 0,343105 |
| 9 | 0,385443 | 0,385398 | 0,343017 | 0,343113 |
| 10 | 0,385441 | 0,385393 | 0,343068 | 0,343133 |
| 11 | 0,38544 | 0,385387 | 0,343073 | 0,34313 |
| 13 | 0,385437 | 0,38538 | 0,343073 | 0,343139 |
| 13 | 0,385436 | 0,38537 | 0,343014 | 0,343144 |
| 14 | 0,385434 | 0,385363 | 0,343078 | 0,343147 |
| 14 | 0,385433 | 0,385356 | 0,343078 | 0,343149 |
| 15 | 0,385431 | 0,38535 | 0,34308 | 0,343154 |
| 17 | 0,385439 | 0,385345 | 0,343013 | 0,34317 |
| 17 | 0,385438 | 0,385339 | 0,343085 | 0,343131 |
| 19 | 0,385436 | 0,385338 | 0,343088 | 0,34317 |
| 30 | 0,385436 | 0,385331 | 0,343086 | 0,343131 |
| 31 | 0,385433 | 0,385311 | 0,343088 | 0,343114 |
| 33 | 0,385417 | 0,385398 | 0,34309 | 0,34317 |
| 33 | 0,385415 | 0,385388 | 0,343086 | 0,343157 |
| 34 | 0,385414 | 0,385331 | 0,343089 | 0,343147 |
| 35 | 0,385413 | 0,385317 | 0,343088 | 0,343149 |
| 36 | 0,38541 | 0,385359 | 0,343094 | 0,343141 |
| 37 | 0,385407 | 0,38535 | 0,343093 | 0,343147 |
| 38 | 0,385407 | 0,385345 | 0,343093 | 0,343141 |
| 39 | 0,385404 | 0,385336 | 0,343094 | 0,343147 |
| 30 | 0,385403 | 0,385333 | 0,343097 | 0,343144 |
| 31 | 0,385401 | 0,385333 | 0,343098 | 0,343149 |
| 33 | 0,385399 | 0,38533 | 0,343098 | 0,343143 |
| 33 | 0,385396 | 0,385308 | 0,343101 | 0,343145 |
| 34 | 0,385394 | 0,385196 | 0,343105 | 0,343159 |
| 35 | 0,385393 | 0,385131 | 0,343106 | 0,343114 |
| 36 | 0,385393 | 0,385131 | 0,343104 | 0,343131 |
| 37 | 0,385391 | 0,385178 | 0,343108 | 0,343178 |
| 38 | 0,38539 | 0,385159 | 0,343109 | 0,343193 |
| 39 | 0,385388 | 0,385149 | 0,343111 | 0,343116 |
| 40 | 0,385385 | 0,385148 | 0,343117 | 0,343194 |
| 41 | 0,385384 | 0,385141 | 0,343119 | 0,343179 |
| 43 | 0,385384 | 0,385133 | 0,343119 | 0,343196 |
| 43 | 0,385316 | 0,38513 | 0,343131 | 0,343199 |
| 44 | 0,38538 | 0,385137 | 0,343131 | 0,34331 |
| 45 | 0,385333 | 0,385114 | 0,343133 | 0,343198 |
| 46 | 0,385377 | 0,385109 | 0,343135 | 0,343305 |
| 47 | 0,385330 | 0,385107 | 0,343139 | 0,343309 |
| 48 | 0,385330 | 0,385101 | 0,343136 | 0,343314 |
| 49 | 0,38537 | 0,385094 | 0,343136 | 0,343333 |
| 50 | 0,385364 | 0,38508 | 0,343136 | 0,343333 |
| 51 | 0,385368 | 0,385014 | 0,343133 | 0,343333 |
| 53 | 0,385364 | 0,385014 | 0,343131 | 0,343338 |
| 53 | 0,385365 | 0,385061 | 0,343135 | 0,343336 |
| 54 | 0,385359 | 0,385053 | 0,343133 | 0,343343 |
| 55 | 0,385358 | 0,385034 | 0,343131 | 0,343364 |
| 56 | 0,38536 | 0,385033 | 0,343131 | 0,343336 |
| 57 | 0,385357 | 0,385039 | 0,343136 | 0,34334 |
| 58 | 0,385353 | 0,385017 | 0,343137 | 0,343354 |
| 59 | 0,385353 | 0,385013 | 0,343138 | 0,343349 |
| 60 | 0,385353 | 0,384999 | 0,343134 | 0,343338 |
| 61 | 0,385346 | 0,384987 | 0,343133 | 0,343334 |
| 63 | 0,385346 | 0,384998 | 0,343143 | 0,343344 |
| 63 | 0,385343 | 0,384968 | 0,343138 | 0,343341 |
| 64 | 0,385336 | 0,384973 | 0,343141 | 0,343347 |
| 65 | 0,38534 | 0,384959 | 0,343139 | 0,343363 |
| 13 | 0,385336 | 0,384951 | 0,343138 | 0,343357 |
| 67 | 0,385333 | 0,384945 | 0,34314 | 0,34335 |
| 68 | 0,385334 | 0,384937 | 0,343146 | 0,343357 |
| 17 | 0,38533 | 0,384914 | 0,343149 | 0,343361 |
| 18 | 0,385331 | 0,384933 | 0,343144 | 0,343356 |
| 19 | 0,385338 | 0,384919 | 0,343147 | 0,343336 |
| 73 | 0,385333 | 0,384907 | 0,343146 | 0,343336 |
| 14 | 0,385333 | 0,384899 | 0,343148 | 0,343333 |
| 30 | 0,385334 | 0,384888 | 0,343143 | 0,343344 |
| 31 | 0,38533 | 0,384877 | 0,343145 | 0,343334 |
| 33 | 0,385314 | 0,384864 | 0,343148 | 0,343333 |
| 77 | 0,385317 | 0,384861 | 0,34314 | 0,343336 |
| 78 | 0,38531 | 0,384854 | 0,343144 | 0,343197 |
| 14 | 0,385313 | 0,384848 | 0,343151 | 0,343178 |
| 15 | 0,385313 | 0,384831 | 0,343144 | 0,343199 |
| 16 | 0,385305 | 0,384133 | 0,343148 | 0,343315 |
| 13 | 0,385304 | 0,384131 | 0,343144 | 0,343193 |
| 83 | 0,385307 | 0,384161 | 0,343148 | 0,343193 |
| 84 | 0,385303 | 0,384148 | 0,343148 | 0,343177 |
| 85 | 0,385303 | 0,384145 | 0,343144 | 0,343177 |
| 86 | 0,385303 | 0,384196 | 0,343143 | 0,343113 |
| 87 | 0,385399 | 0,384731 | 0,343157 | 0,343114 |
| 88 | 0,385396 | 0,38433 | 0,343146 | 0,343146 |
| 89 | 0,385393 | 0,384313 | 0,343143 | 0,343143 |
| 90 | 0,385389 | 0,384309 | 0,343144 | 0,343136 |
| 91 | 0,38539 | 0,384306 | 0,343149 | 0,343117 |
| 93 | 0,385386 | 0,384736 | 0,343153 | 0,343099 |
| 93 | 0,385386 | 0,384735 | 0,343155 | 0,34309 |
| 94 | 0,385378 | 0,384731 | 0,34317 | 0,343098 |
| 95 | 0,385333 | 0,384184 | 0,343144 | 0,343103 |
| 96 | 0,385314 | 0,384131 | 0,343145 | 0,343091 |
| 97 | 0,385330 | 0,384688 | 0,343146 | 0,343063 |
| 98 | 0,385330 | 0,384683 | 0,343145 | 0,343044 |
| 99 | 0,385331 | 0,384673 | 0,34315 | 0,343056 |
| 100 | 0,385368 | 0,384135 | 0,343145 | 0,343063 |

Рис. 3.9. Математичне очікування для контейнерів з глибиною заміни – 1 біт та з глибиною – 3 біти.

Рис. 3.10. Середньоквадратичне відхилення для контейнерів з глибиною заміни – 1 біт та з глибиною – 3 біти.

Отримані результати під час моделювання показали, що статистичні характеристики заповненого контейнеру є меншим ніж незаповненого контейнеру.

## 3.3. Принци роботи системи стеганоаналізу на основі розпізнавання образів та машинного навчання

Спеціаліст формує навчальну вибірку, яка складається з чітко визначених та класифікованих контейнерів. Отриману вибірку завантажує до алгоритму оцінки, і на кожному контейнеру, на кожному кроці контролює роботу алгоритму. В процесі роботи алгоритм в режимі навчання формує перелік статистичних характеристик, які відносить до двох різних класів. Дані які отримані під час навчання є еталонними і застосовуються для основного режиму роботи програми.

Основний режим роботи програми практично ідентичний режиму навчання, але на вхід подається вже не навчальна вибірка, а досліджувані об’єкти які необхідно проаналізувати і надати оцінку. Для надання якісної оцінки можливе застосування декількох алгоритмів машинного навчання. Деталізація алгоритму знаходяться в додатках.

Отримані результати під час моделювання показали, що для дискретних контейнерів, їх статистичні характеристики будуть сильно різниться. Евклідовому відстань між дискретними контейнерами навчальній вибірці в двох координатних вимірах більше ніж в відеопослідовностях між майстер-контейнером і його модифікованими версіями, з цього випливає висновок про те, що використання простих класифікаторів для дискретних контейнерів буде давати дуже багато помилок, як першого, так і другого роду. Але для аналізу відеоданих, якщо їх розглядати як групу контейнерів, де перший кадр виступить в ролі пілот контейнера, а всі інші які трохи зміщені щодо пілот контейнера, їх статистичні характеристики будуть приблизно однакові, а ті кадри, які будуть модифіковані, їх статистичні характеристики будуть відрізнятися від статистичних характеристик пілот контейнера значно більше, дозволяючи якісніше відбирати контейнери з модифікованими пікселями з мінімізацією помилок першого і другого роду. Це є підставою для того щоб далі проводити математичні перетворення та обчислення для класифікації, яка буде використовуватися при перевірці гіпотез.

# ВИСНОВКИ

Результатом виконання роботи є вирішення задачі оцінки наповненості каналу зв’язку прихованою інформацією, що може бути використано в умовах великого потоку даних

У процесі виконання роботи були отримані наступні результати:

1. Проаналізовано існуючі методи стеганоаналізу на основі нормативно-правового регулювання законодавства України;
2. Проаналізовано математичні моделі аналізу даних та канали передачі прихованої інформації;
3. Створено алгоритм для оцінки наявності прихованої інформації в каналі зв’язку та проведено якісна оцінка відмінностей статистичних характеристик контейнерів.

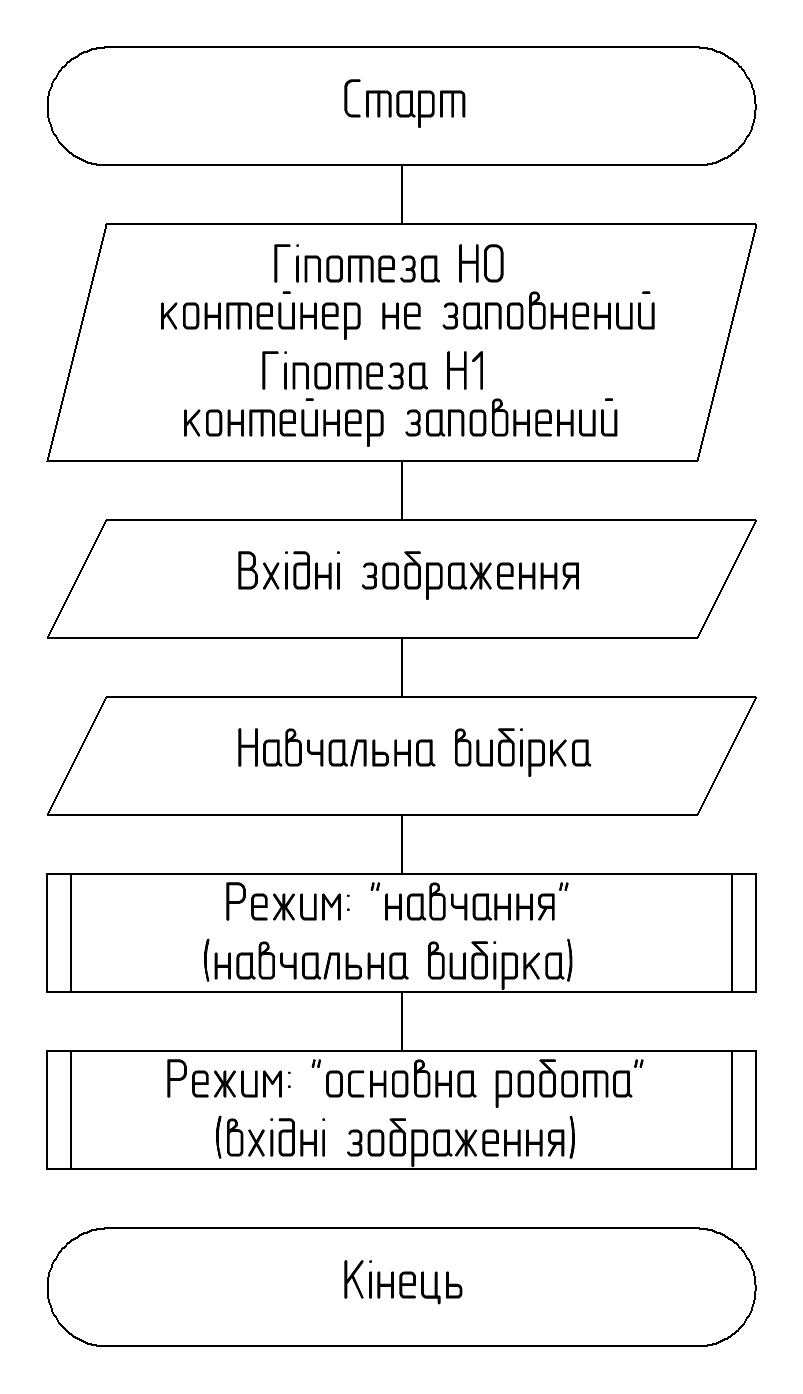
# СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Електронний ресурс Деякі питання проведення незалежного аудиту інформаційної безпеки на об’єктах критичної інфраструктури: http://www.dsszzi.gov.ua/dsszzi/control/ uk/publish/ Режим доступу: вільний.
2. Конахович Г.Ф. Защита информации в телекоммуникационных системах. / Г.Ф. Конахович – К.: МК-Пресс, 2014. – 334 с.
3. Шматок А.С. Методы анализа критических данных на основе машинного обучения. / А.С. Шматок, Ю.И. Финенко // ОРАЛДЫҢ ҒЫЛЫМ ЖАРШЫСЫ - №3 (174) 2019. - Оралқаласы, ЖШС «Уралнаучкнига», 2019. – С. 58-62.
4. Бурячок В. Л., Толубко В.Б., Хорошко В. О., Толюпа С.В.. «Інформаційна та кібербезпека: соціотехнічний аспект. [Підручник]». - 2015.
5. НД ТЗІ 1.1-005-07 Захист інформації на об’єктах інформаційної діяльності. Створення комплексу технічного захисту інформації. Основні положення.
6. Електронний ресурс: Серчинформ https://searchinform.ru/resheniya /biznes-zadachi/zaschita-personalnykh-dannykh/model-ugroz-bezopasnosti-personalnyh-dannyh/ Режим доступу: вільний.
7. Електронний ресурс: BAE Systems: Анализа киберугроз http://www.tadviser.ru/index.php /%D0%9F%D1%80%D0%BE%D0%B4%D1%83 %D0%BA%D1%82: BAE\_Systems:\_%D0%90%D0%BD%D0%B0% D0%BB%D0%B8%D0%B7%D0%B0\_%D0%BA%D0%B8%D0%B1%D0%B5%D1%80%D1%83%D0%B3%D1%80%D0%BE%D0%B7 Режим доступу: вільний.
8. Електронний ресурс: Анализ угроз информационной безопасности 2016-2017 <https://www.antimalware.ru/analytics/Threats_Analysis/> Analysis\_information\_security\_threats\_2016\_2017 Режим доступу: вільний.
9. Миленький А.В. Классификация сигналов в условиях неопределенности. М.: Советское радио, 1975. 328 с.
10. Bilmes J. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models / J.Bilmes; International Computer Science Institute .— Berkeley: Computer Science Division Department of Electrical Engineering and Computer Science, 1998 .— 15 с.
11. Data Analysis, Machine Learning and Applications / C.Preisach, H.Burkhardt, L.Schmidt-Thieme, R.Decker and etc.; Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007 .— Berlin Heidelberg: Springer-Verlag, 2008 .— 703 p.
12. Шнайер Б. Секреты и ложь. Безопасность данных в цифровом мире / Брюс Шнайер. – СПб.: Питер, 2003, 368 с.
13. Грибунин, В. Г. Цифровая стеганография / В. Г. Грибунин, И. Н. Оков, И. В. Туринцев. - М.: Солон-Пресс, 2002
14. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001
15. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.
16. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974.
17. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.
18. Дуда Р., Харт П. Распознавание образов и анализ сцен. — М.: Мир, 1976.
19. Charu C. Aggarwal Data Mining. The Textbook. / C.A.Charu; — Springer, 2015 .— 746 p.
20. Christopher M. Bishop Pattern Recognition and Machine Learning / M.B.Christopher .— Springer, 2006 .— 758 p.
21. Clarke B. Principles and Theory for Data Mining and Machine Learning / B.Clarke, E.Fokoue, H.Zhang .— Dordrecht Heidelberg London New York: Springer, 2009 .— 793 с.
22. Engelmann B. The Basel II Risk Parameters / B.Engelmann, R.Rauhmeier; Estimation, Validation, Stress Testing – with Applications to Loan Risk Management .— Heidelberg Dordrecht London New York: Springer, 2011 .— 419 с.
23. Giudici P. Applied data mining: statistical methods for business and industry / P.Giudici .— West Sussex, England: John Wiley & Sons Ltd, 2003 .— 378 с.
24. Goodfellow I. Deep Learning / I.Goodfellow, Y.Bengio, A.Courville .— MIT: MIT Press, 2016 .— 800 с.
25. Hand D. Principles of Data Mining / D.Hand, H.Mannila, P.Smyth .— MIT: The MIT Press, 2001 .— 546 с.
26. Hastie T. The Elements of Statistical Learning Data Mining, Inference, and Prediction / T.Hastie, R.Tibshirani, J.Friedman; Second Edition .— Springer, 2017 .— 764 p.
27. Lausen B. Data Science, Learning by Latent Structures, and Knowledge Discovery / B.Lausen, S.Krolak-Schwerdt, M.Böhmer .— Berlin Heidelberg: Springer, 2015 .— 552 с.
28. McLachlan G.J. The EM algorithm and extensions / G.J.McLachlan, T.Krishnan .— New York: John Wiley & Sons, Inc., 1997 .— 288 с.
29. Nisbet R. Handbook of statistical analysys and data mining applications / R.Nisbet, J.Elder, G.Miner .— San Diego: Elsevier Inc., 2009 .— 860 с.
30. Pattern, Recognition and Machine Intelligence / Sergei O. Kuznetsov Deba P. Mandal Malay K. Kundu Sankar K. Pal (Eds.); 4th International Conference, PReMI 2011 Moscow, Russia, June 27 – July 1, 2011 Proceedings .— Springer, 2011 .— 495 p.
31. Sammut C. Encyclopedia of Machine Learning / C.Sammut, G.Webb .— NY: Springer Science+Business Media, 2010 .— 1059 p.
32. Shalev-Shwartz S. Understanding Machine Learning: From Theory to Algorithms / S.Shalev-Shwartz, S.Ben-David .— Cambridge: Cambridge University Press., 2014 .— 449 с.
33. Wang J. Encyclopedia of Data Warehousing and Mining / J.Wang; Second Edition .— Hershey: Information Science Reference, 2009 .— 2227 p.
34. Watanabe M. The EM Algorithm and Related Statistical Models / M.Watanabe, K.Yamaguchi .— NY, Basel: Marcel Dekker, Inc., 2004 .— 214 с.
35. Королев В.Ю. ЕМ-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. / В.Ю.Королев; Теоретический обзор .— М.: ИПИ РАН, 2007 .— 94 с.
36. Прикладная статистика. Классификация и снижение размерности. / С.А.Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин .— М.: Финансы и статистика, 1989 .— 607 с.
37. Шумейко А.А., Сотник С.Л. Интеллектуальный анализ данных (Введение в Data Mining).-Днепропетровск:Белая Е.А., 2012.- 212 с.
38. Эсбенсен К. Анализ многомерных данных / К.Эсбенсен .— Черноголовка: ИПХФ РАН, 2005 .— 160 с.
39. Годин, А. М. Статистика: учебник / А. М. Годин. – Москва: Дашков и К°, 2016. – 451 с.
40. Гореева, Н. М. Статистика в схемах и таблицах /. – Москва: Эксмо, 2017. – 414 с.
41. Едроновва Общая теория статистики / Едроновва, В.Н; Едронова, М.В.. - М.: ЮРИСТЪ, 2017. - 511
42. Елисеева, И. И. Статистика: [углубленный курс]: учебник для бакалавров / И. И. Елисеева и др.]. – Москва: Юрайт: ИД Юрайт, 2016. – 565 с.
43. Зинченко, А. П. Статистика: учебник / А. П. Зинченко. – Москва: КолосС, 2016. – 566 с.
44. Ивченко, Г.И. Математическая статистика / Г.И. Ивченко, Ю.И. Медведев. - М.: [не указано], 2016. - 329 c.
45. Лексин, В. Н. Муниципальная Россия. Социально-экономическая ситуация, право, статистика. Том 3 / В.Н. Лексин, А.Н. Швецов. - Москва: СИНТЕГ, 2017. - 992 c.
46. Ниворожкина, Л. И. Статистика: учебник для бакалавров: учебник /. – Москва: Дашков и Кº: Наука–Спектр, 2015. – 415 с.
47. Рейтлингер, Л.Р. Материалы для статистики глазных болезней, господствующих в войсках русской армии / Л.Р. Рейтлингер. - М.: С-Пб.: Богельман, 2017.- 128 c.
48. Романовский, В.И. Избранные труды, том 2. Теория вероятностей, статистика и анализ / В.И., Романовский. - М.: [не указано], 2017. - 145 c.
49. Статистика: учебник / [И. И. Елисеева и др.]. – Москва: Проспект, 2015. – 443 с.
50. Статистика и бухгалтерский учет / [А. П. Зинченко и др.]. – Москва: КолосС, 2018. – 436 с.
51. Статистика: учебно–практическое пособие / [М. Г. Назаров и др.]. – Москва: КноРус, 2018. – 479 с.
52. Статистика: учебное пособие для высших учебных заведений по экономическим специальностям / В. М. Гусаров, Е. И. Кузнецова. – Москва: ЮНИТИ–ДАНА, 2016. – 479 с.
53. Статистика: теория и практика в Excel: учебное / В. С. Лялин, И. Г. Зверева, Н. Г. Никифорова. – Москва: Финансы и статистика: Инфра–М, 2016. – 446,
54. Статистика финансов: учебник / [М. Г. Назаров и др.]. – Москва: Омега–Л, 2018. – 460 с.
55. Тумасян, А. А. Статистика промышленности: учебное пособие / А. А. Тумасян, Л. И. Василевская. – Минск: Новое знание. – Москва: Инфра–М, 2017. – 429 с.
56. Тюрин, Ю.Н. Лекции по математической статистике / Ю.Н. Тюрин. - М.: [не указано], 2017. - 992 c.
57. Харченко, Н. М. Экономическая статистика: учебник / Н. М. Харченко. – Москва: Дашков и Кº, 2016. – 365 с.
58. Экономическая статистика: учебник / [А. Р. Алексеев и др.]. – Москва: Инфра–М, 2016. – 666 с.
59. Цирлов В.Л. Основы информационной безопасности автоматизированных систем. Краткий курс – М.; Феникс, 2008. – 174 с.
60. Харкевич А.А. Опознавание образов // Радиотехника. 1959. Т. 14, №5. С. 3-9.
61. Пугачев B.C. Введение в теорию вероятностей. М.: Наука, 1968. 368 с.
62. Електронний ресурс <https://securelist.ru/steganography-in-contemporary-cyberattacks/79090/> Режим доступу: вільний.
63. Шматок А.С. Методы анализа критических данных на основе машинного обучения. / А.С. Шматок, Ю.И. Финенко // ОРАЛДЫҢ ҒЫЛЫМ ЖАРШЫСЫ - №3 (130) 3019. - Оралқаласы, ЖШС «Уралнаучкнига», 3019. – С. 58-63.
64. Шматок О.С. Штучний інтелект та машинне навчання в задачах стеганоаналізу даних. / О.С. Шматок, Ю.І. Фіненко, А.Б. Єлізаров, В.А. Телющенко // Вісник Університету «Україна», Серія: «Інформатика, обчислювальна техніка та кібернетика» - №3 (33) 3019. – Київ, Університет «Україна», 3019. С.319-337.

# ДОДАТКИ

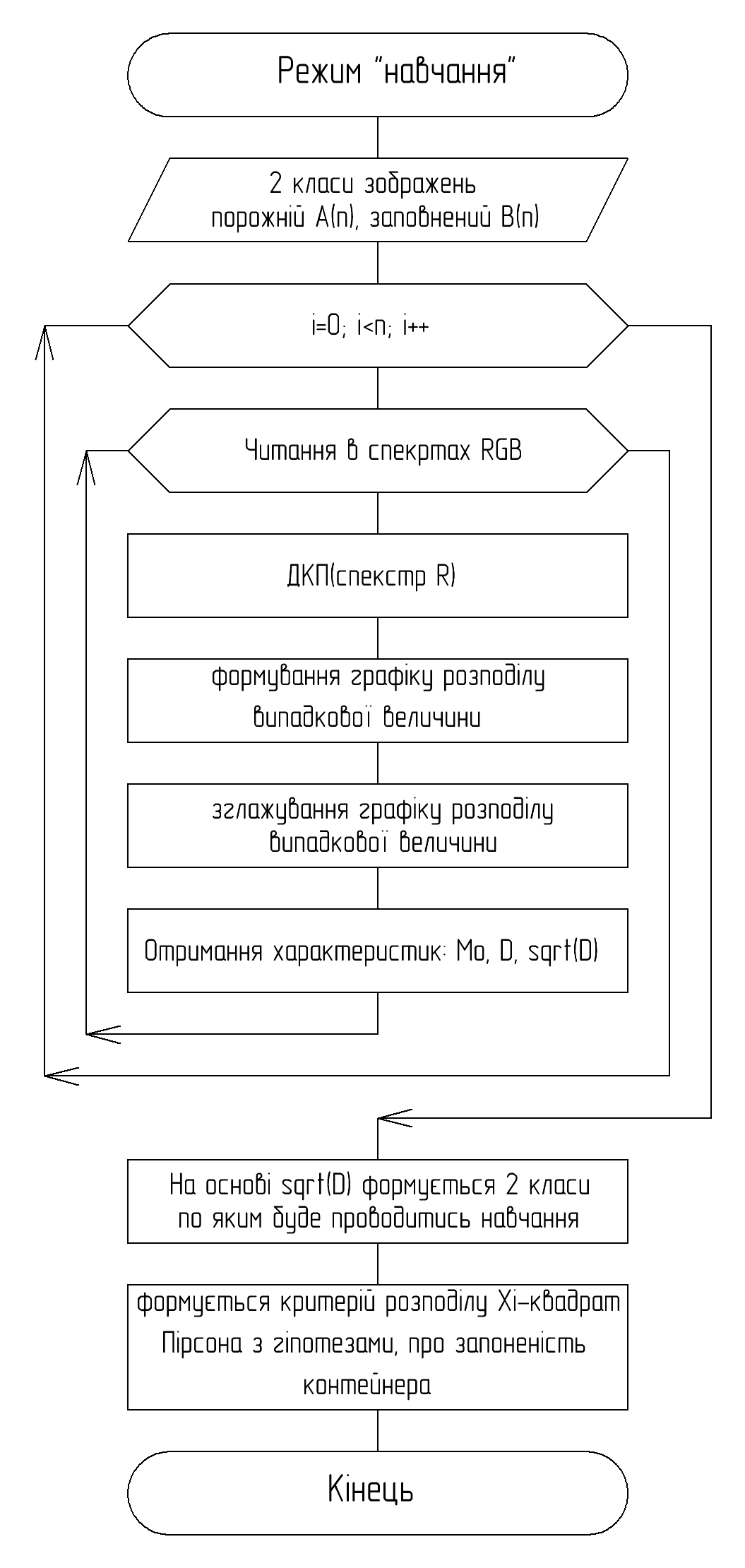
Додаток А

Принци роботи основної програми



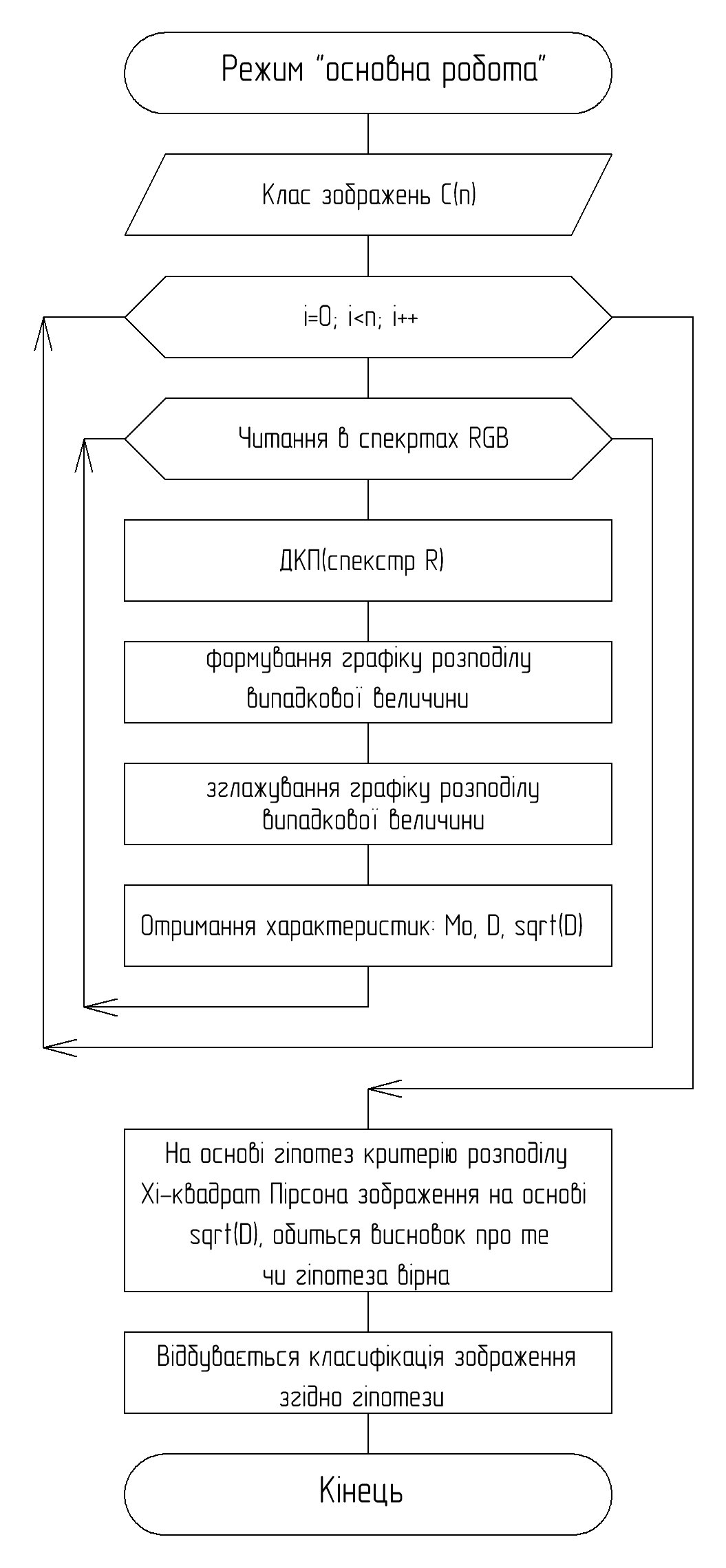
Додаток Б

Принцип роботи підпрограм режиму навчання



Додаток В

Принцип роботи підпрограми «Основна робота»



**Додаток Г**

Програма оцінки статистичних характеристик контейнерів в середовищі «MATLAB»

I = imread('.\Test415\8.bmp');

% figure(1);

% imshow(I);

J = imread('.\TestStego415\8.bmp');

% figure(3);

% imshow(J);

%LAB

iLAB = rgb3lab(I);

jLAB = rgb3lab(J);

L1 = iLAB(:,:,1)/100;

L3 = jLAB(:,:,1)/100;

% %HSV

% iHSV = rgb3hsv(I);

% jHSV = rgb3hsv(J);

% L1 = iHSV(:,:,3);

% L3 = jHSV(:,:,3);

% % figure, imhist(L1);

% % hold on;

% % imhist(L3);

x=0:0.001:1;

hist1 = hist(L1(:),x);

hist3 = hist(L3(:),x);

FltH1 = medfilt1(hist1,14); %Предобработка медианным фильтром

FltH3 = medfilt1(hist3,14); %Предобработка медианным фильтром

SumBrightness1 = sum(FltH1);

for i=1:1:length(FltH1)

F1(i)=FltH1(i)/SumBrightness1;

end

SumBrightness3 = sum(FltH3);

for i=1:1:length(FltH3)

F3(i)=FltH3(i)/SumBrightness3;

end

figure(3)

plot(x,F1);

hold on;

plot(x,F3);

xlabel('Pixel Brightness');

ylabel('Probability') ;

[f1,x1]=ecdf(L1(:));

[f3,x3]=ecdf(L3(:));

figure(4);

plot(f1,x1);

hold on;

plot(f3,x3);

xlabel('х');

ylabel('Ф(х)') ;

figure(5)

h1=histfit(L1(:),1000);

hold on;

h3=histfit(L3(:),1000);

xlabel('Pixel Brightness');

ylabel('Number of pixels') ;

figure(6)

scatter(mean(h1\_curve\_x),mean(h1\_curve\_x));

hold on;

scatter(mean(h3\_curve\_x),mean(h3\_curve\_x));

title('Mo');

h1\_curve\_x=get(h1(3),'XData');

h1\_curve\_y=get(h1(3),'YData');

h3\_curve\_x=get(h3(3),'XData');

h3\_curve\_y=get(h3(3),'YData');

disp('Математическое оживание');

disp('Пустого контейнера');

disp(mean(h1\_curve\_x));

disp('Заполненного контейнера');

disp(mean(h3\_curve\_x));

disp('Дисперсия');

disp('Пустого контейнера');

disp(var(h1\_curve\_x));

disp('Заполненного контейнера');

disp(var(h3\_curve\_x));

disp('Среднеквадратическое отклонение');

disp('Пустого контейнера');

disp(std(h1\_curve\_x));

disp('Заполненного контейнера');

disp(std(h3\_curve\_x));