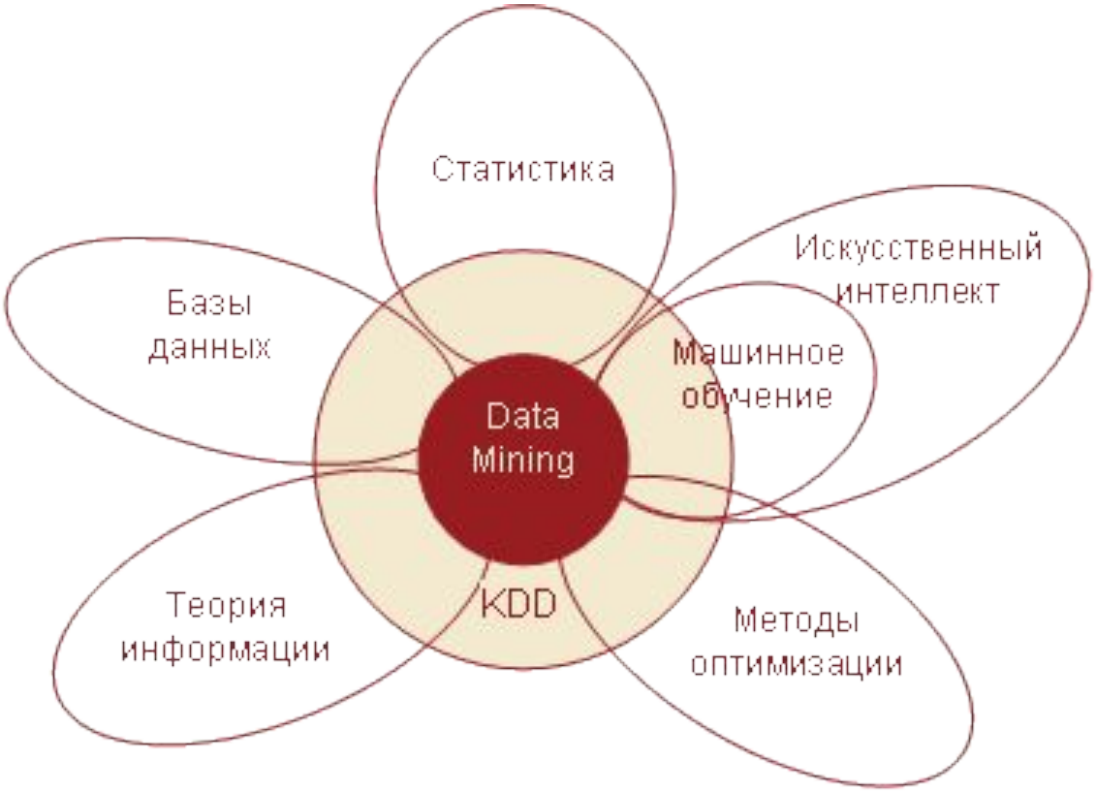
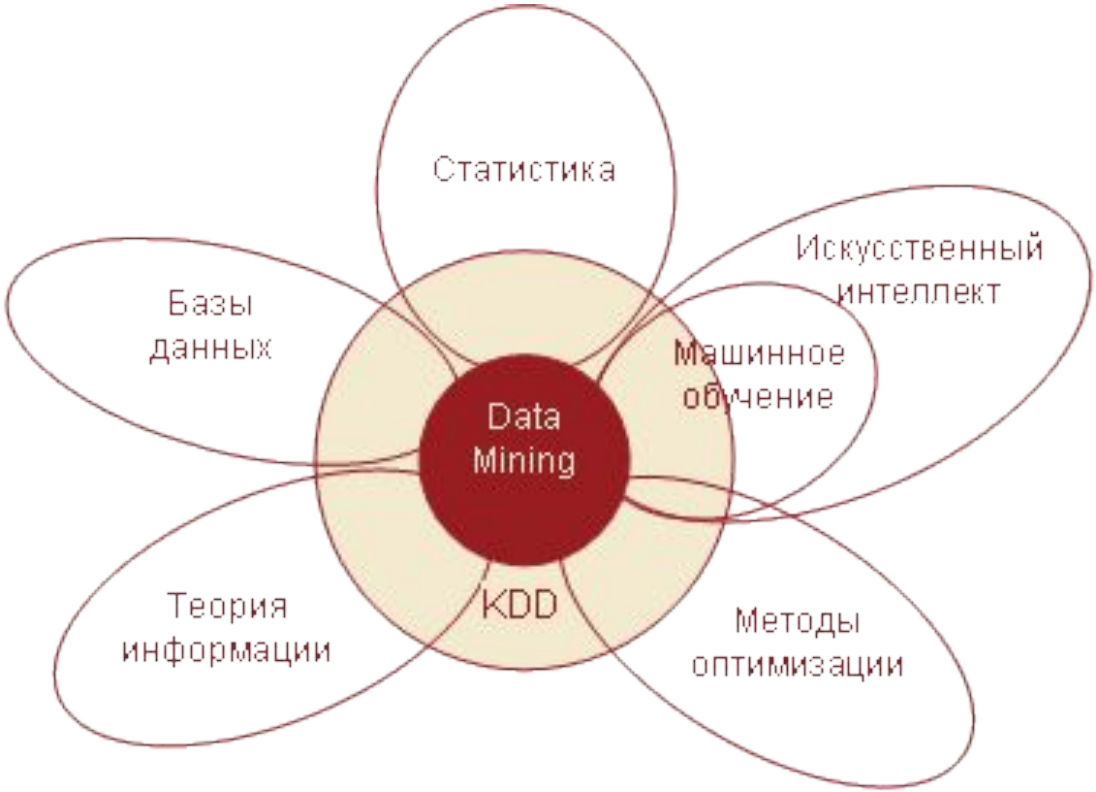
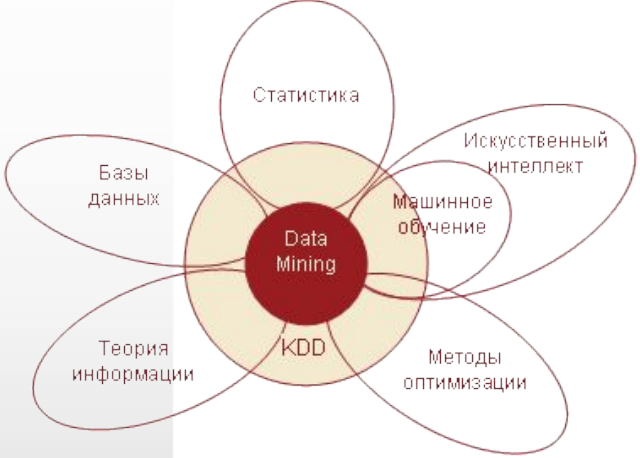
Загальні поняття про інтелектуальний аналіз даних



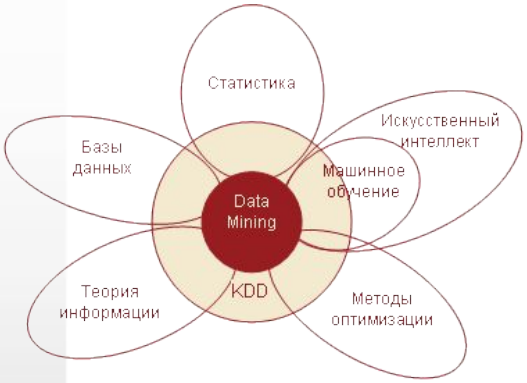
Лекція1«ШАМАНСТВО» В АНАЛіЗі ДАНИХ





План

1. Передмова
2. Управління силою думки
3. Дивовижні закономірності
4. Правила справжнього шамана



1. Передмова

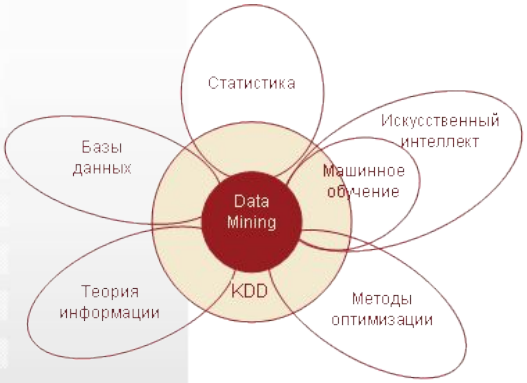
Іноді помітити феномен набагато цінніше, ніж пояснити його.
(Ю.І. Журавльов)

Перша задача цієї лекції - показати, які задачі зустрічаються в аналізі даних.

Аналіз даних - це все, де можна застосувати математику, програмування і, звичайно, здоровий глузд для пошуку закономірностей, інтерпретації даних, прийняття рішення і т.д.

Друга задача - це як раз показати, а що таке здоровий глузд для вирішення завдань аналізу даних.

2. Управління силою думки



Не постачайте дітей готовими формулами, формули - порожнеча, збагатіть їх образами, на яких видно сполучні нитки.
(А.Д. Сент-Екзюпері)

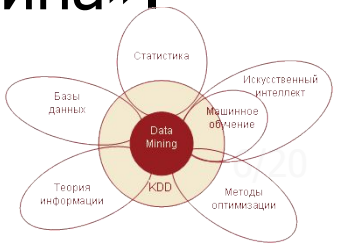
На початку цього століття стали дуже популярними **дослідження в області «Brain Computer Interface »**(Інтерфейс« мозок-комп'ютер »), які якраз
займаються побудовою ефективних інтерфейсів для управління ЕОМ
за допомогою ... сигналів головного мозку.

Людина сідає перед комп'ютером, а йому на голову одягається шапочка з електродами, яка підключається до комп'ютера.

Під час ментальних дій (по-простому «роздумів») змінюється потенційне і магнітне поле різних ділянок головного мозку, все це фіксується за допомогою пристосування шапочка-провода-комп'ютер.

Таким чином, комп'ютер знає, що там «відбувається в голові у людини», щоправда, в термінах зміни потенціалу.

Залишилося перевести це на більш зрозумілу мову, щоб комп'ютер розумів, «про що думає людина».



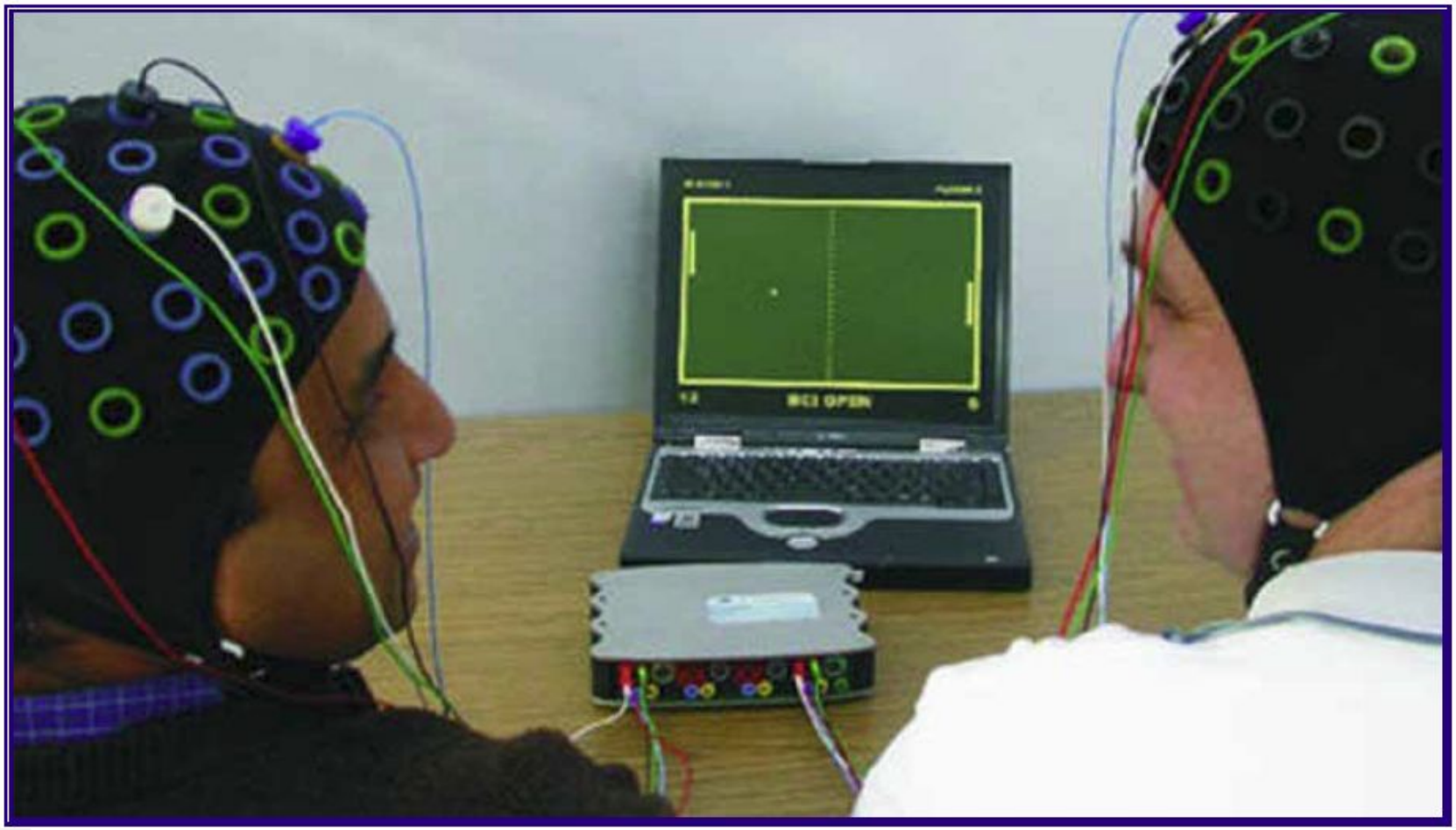
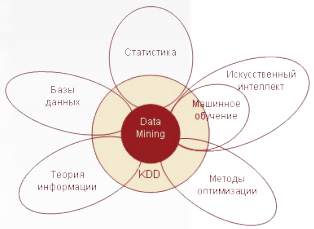
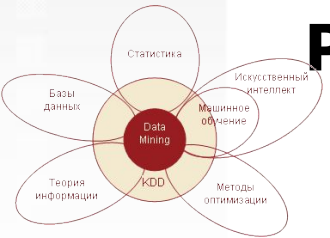


Рис. 1. Гра в теніс за комп'ютером «за допомогою сили думки»



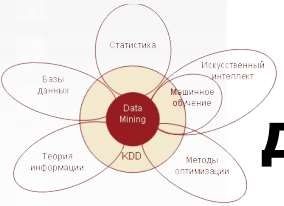
Є спеціальні системи введення слів за допомогою інтерфейсу «мозок-комп'ютер».

Рис.2. Введення тексту в комп'ютер без використання клавіатури



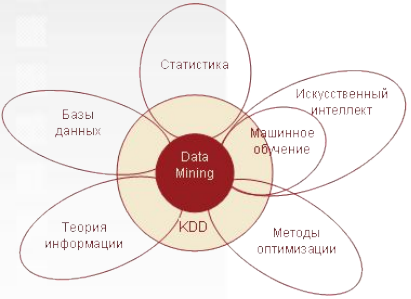
Є спеціальні інвалідні крісла, які приводяться в рух «силою думки».

Рис.3. Управління роботом (протезами) за допомогою інтерфейсу «людина-комп'ютер»



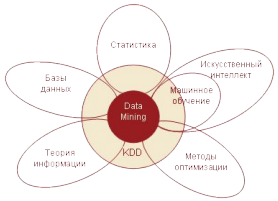
Мова йде не про розуміння думок людини комп'ютером, а про розрізнення декількох ментальних станів (це задача класифікації)

Рис.4. Шапочка з електродами



По тематиці «Brain Computer Interface» було проведено кілька великих міжнародних змагань. Учасникам міжнародного конкурсу з класифікації сигналів «BCI competition III»
2003 була запропонована наступна задача.

Постановка

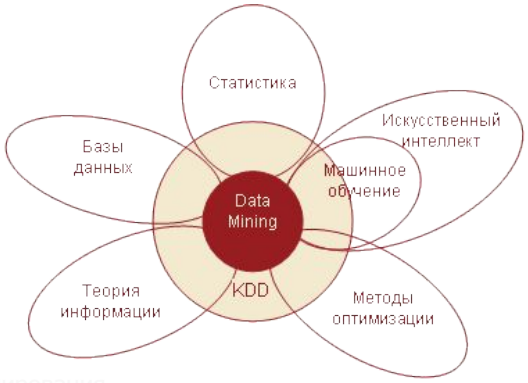


Дано опис 278 сигналів, які відображали два ментальні стани, таким чином були розбиті на два класи. Насправді, це багатовимірні сигнали, оскільки знімались за допомогою ECoG-електродній сітки розміру 8x8 (електродів), тобто одночасно вимірювалося 64 сигнали. Кожен сигнал складався з 3000 точок, оскільки відображав 3-секундну активність головного мозку під час деякої ментальної дії і знімався з частотою 1000Гц. Вихідні дані записуються в тривимірнії матриці розміру 278 × 64 × 3000 (278 64-мірних 3000-точкових сигналу).
Потрібно побудувати алгоритм, який класифікує сигнали. Якість класифікації алгоритму перевірялося на контрольній вибірці з 100 сигналів.

Рішення описаної задачі

Всі ілюстрації, які у нас будуть, відповідають лише одному з 64 сигналів, знятому, в деякому розумінні, з «кращого електрода», який знімав показання з зони мозку, в якій відбувалися «найбільш інтенсивні» зміни.

На (Рис.5.) показано кілька сигналів першого і другого класу, а також один із сигналів, які нам треба класифікувати.



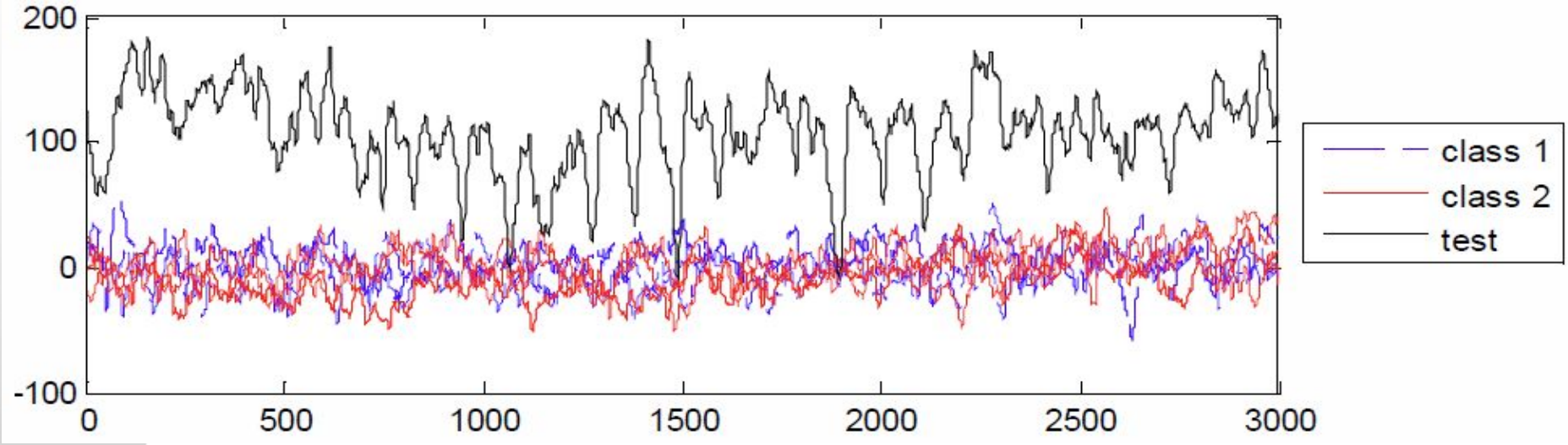
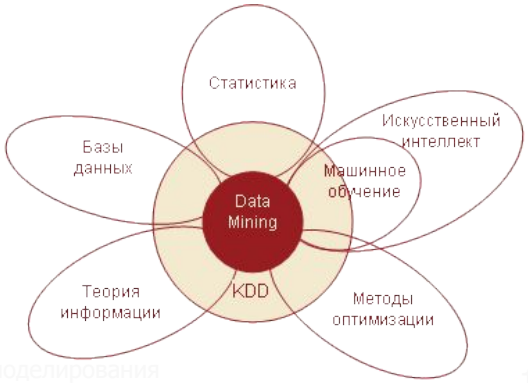
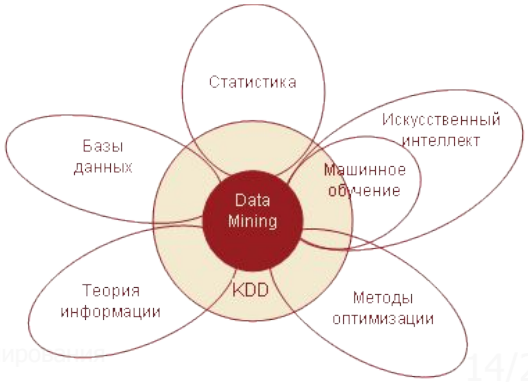


Рис. 5. Візуалізація ECoG-сигналів головного мозку



Найприродніший метод класифікації - подивитися, на сигнали якого класу схожий класифікований(цей метод називається «найближчий сусід»).

Але наші класифіковані сигнали взагалі не схожі на ті, про які відома класифікація! Це наслідок того, що сигнали були отримані в різні дні, тобто зовсім у різних умовах.



Обчислимо для кожного сигналу його мінімальне і максимальне значення. Тоді **сигнал представляється точкою у відповідному просторі.** (Рис.6.)

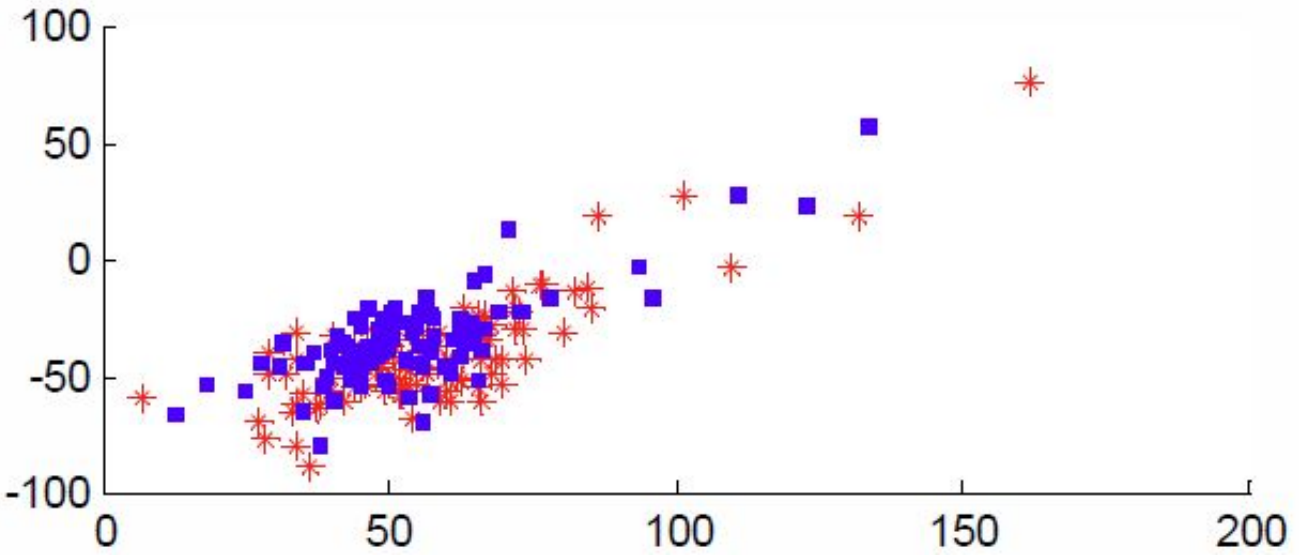
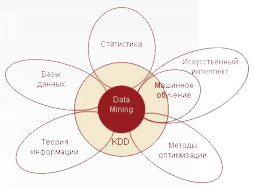
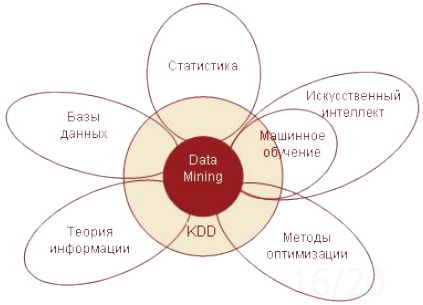


Рис. 6. Максимальні (по горизонталі) і мінімальні (по вертикалі) значення сигналів.



Ще приклад такого простору (рис. 7). Тут показані середні значення сигналів в перші 1,5 секунди (перша ознака) і в останні (друга ознака). Видно, що значення корелюють, тобто за середнім значенням в перші 1,5 секунди «вгадується» значення в останні 1,5 секунди: воно приблизно таке ж.

Чим більше схоже ця хмара точок на лінію, тим точніше ми зможемо «вгадати» друге значення по першому, а якщо хмара точок розмито і на лінію не схоже (рис. 6), то такого вгадування не вийде (ознаки некорельовані, тобто значення одного не визначає значення другого).



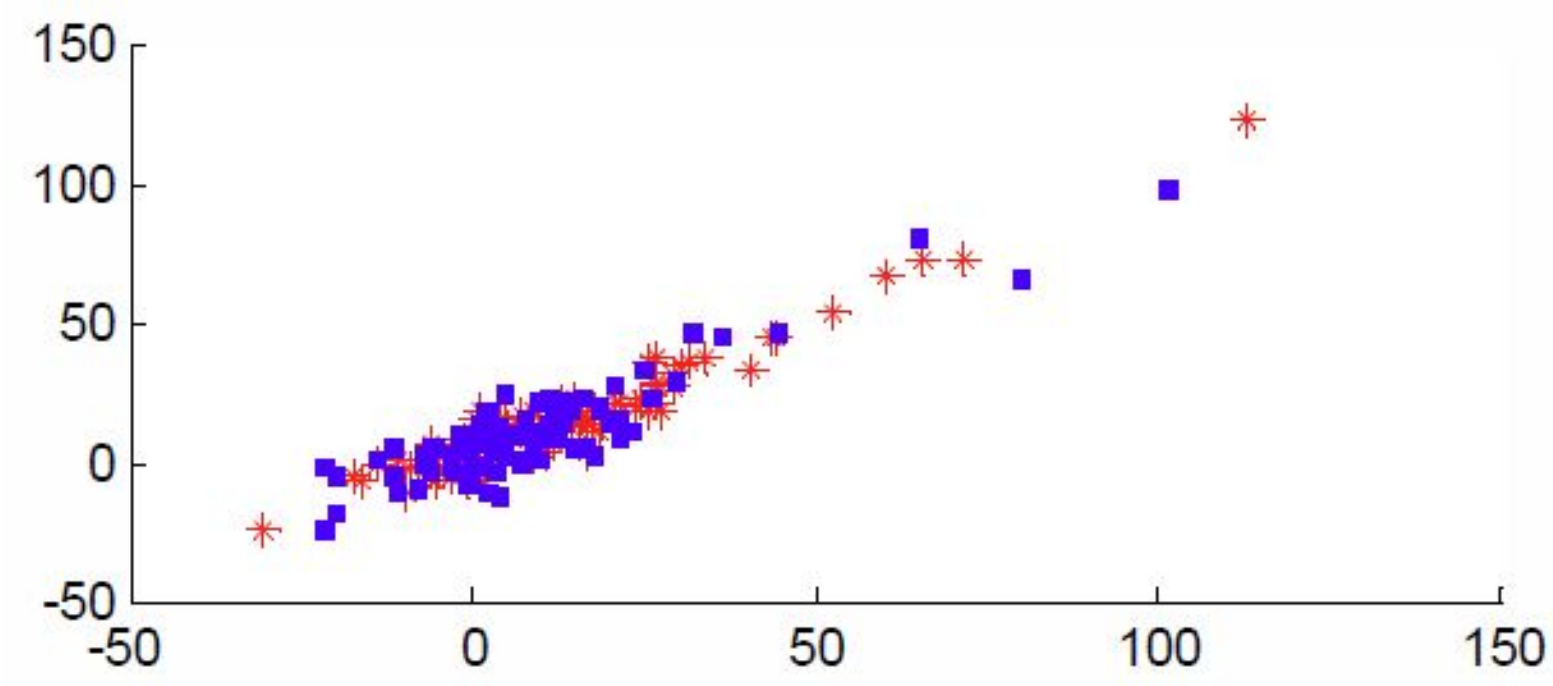
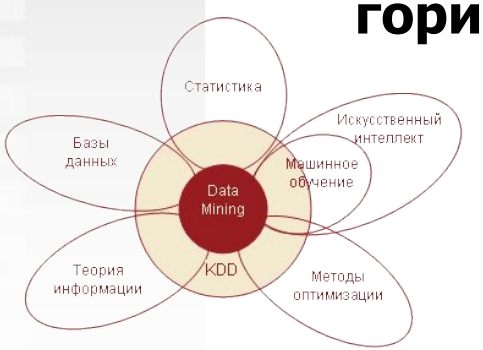


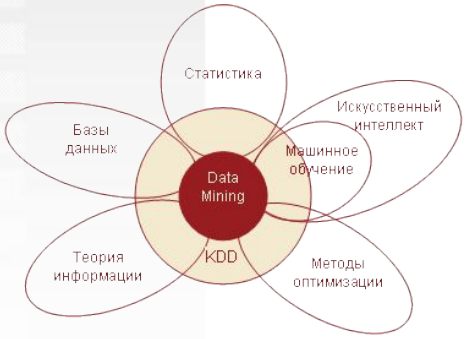
Рис. 7. Середні значення сигналів за перші 1,5 секунди (по горизонталі) і останні 1,5 секунди (по вертикалі).



На (рис.8.) По вертикалі відкладено середнє значення сигналу, а по горизонталі – значення

$$\frac{1}{n-1}\sum_{i=1}^{n-1}|u_{i+1}-u_i|$$

(для сигналу *(u1, ..., un)).* «Фізичний сенс» останньої ознаки зрозуміти неважко: він описує **швидкість зміни сигналу. Дивно, але за цією ознакою сигнали непогано відрізняються:** якщо значення ознаки маленьке, то сигнал, найімовірніше, належить першому класу (синьому), а якщо велика - другому (червоному).



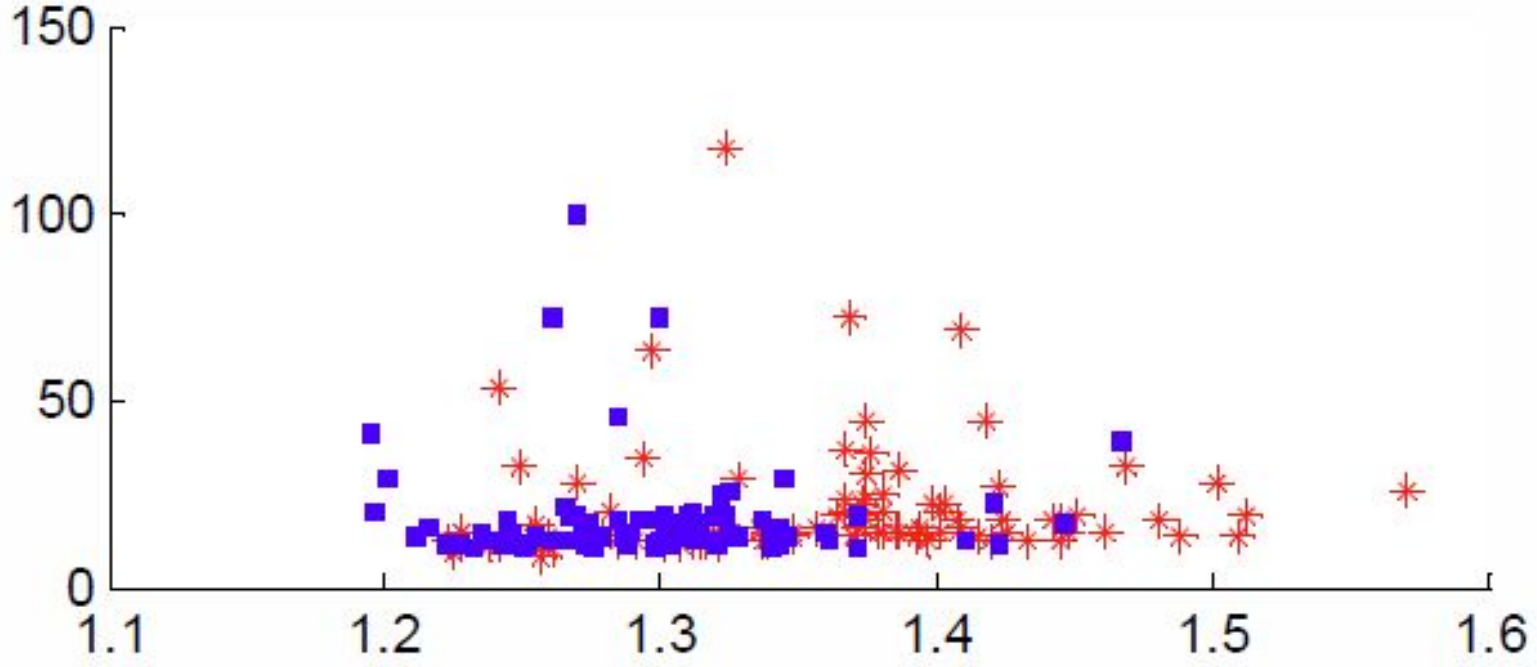
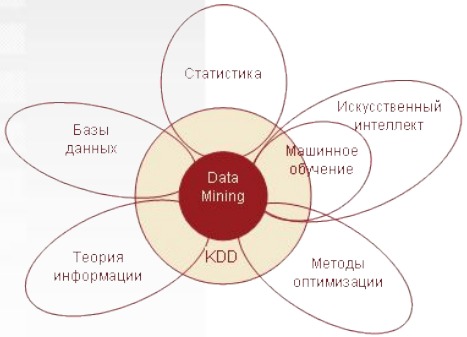


Рис. 8. Середній модуль різниці послідовних значень сигналу (по горизонталі) і середнє значення сигналу (по вертикалі).



Є ще одна ознака, яка також описує «різноманітність значень сигналів»: дисперсія. На рис. 9 зображена пара цих ознак: наш «хороший» і дисперсія.

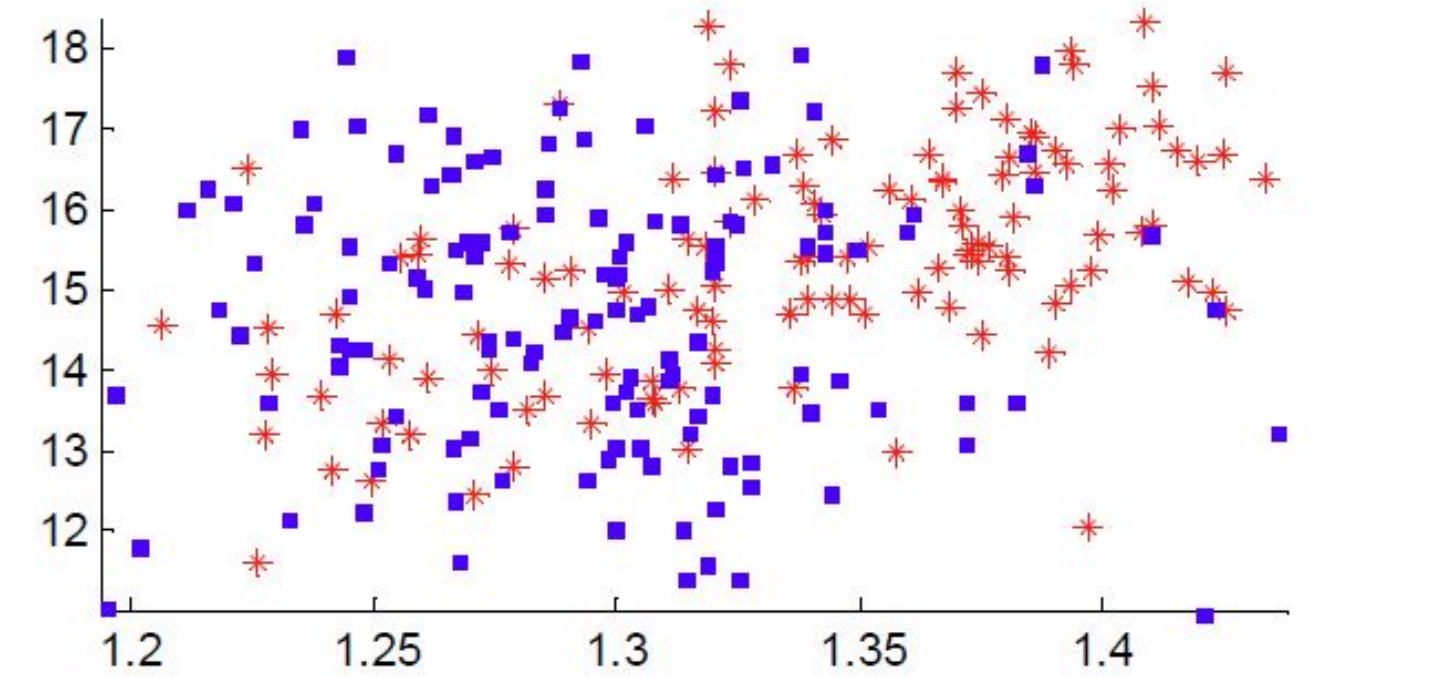
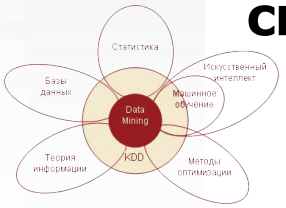


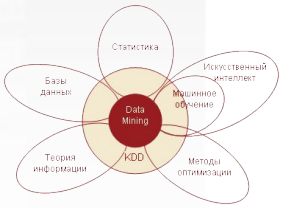
Рис. 9. Середній модуль різниці послідовних значень сигналу (по горизонталі) і дисперсія значень сигналу (по вертикалі).



Ще один стандартний прийом, застосовуваний при візуальному аналізі даних: «трохи» змінити знайдену ознаку. Замість модуля використовувати квадрат:

$$\frac{1}{n-1}\sum_{i=1}^{n-1}(u_{i+1}-u_i)^2$$

Нова ознака, як правило, сильно корелює з вихідною (адже вона отримана її незначною зміною), але в проекції на ці дві ознаки можна побачити цікаві закономірності. На рис. 10 видно невеликий зазор між об'єктами першого і другого класів. Точніше між двома «хмарами точок». У першому хмарі переважають точки першого класу, а в другому - другого.



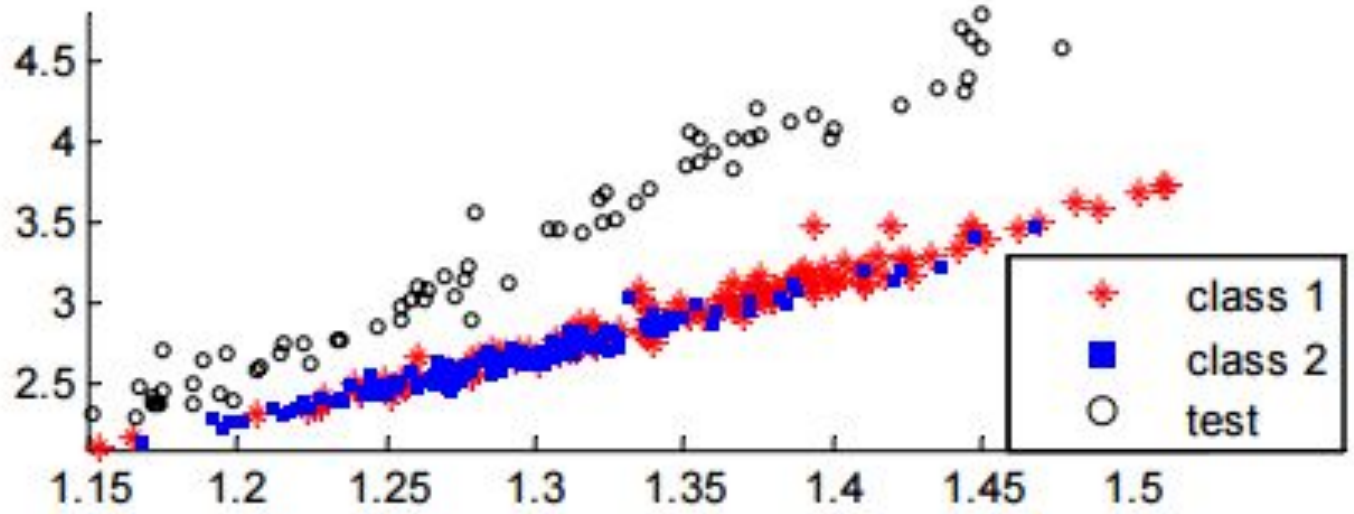
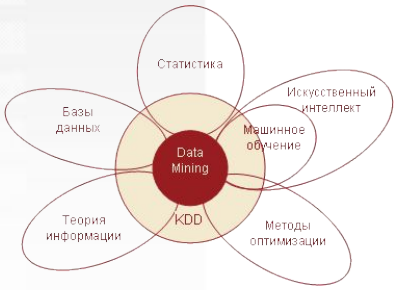
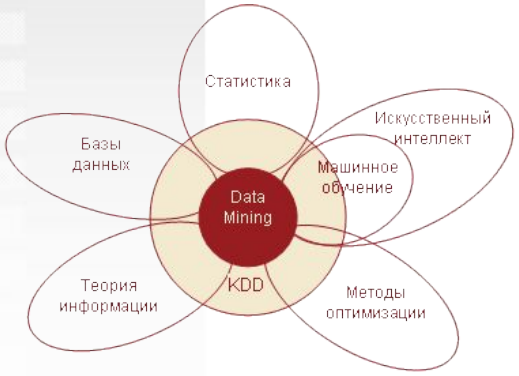


Рис.10. Середній модуль різниці послідовних значень сигналу (по горизонталі) і середній квадрат (по вертикалі).



Отже, ми, власне, вирішили задачу! Звичайно, не з 100% -й точністю (яка тут і не досяжна), але не вдаючись до «високої науки», просто переглядаючи картинки і фантазуючи. Насправді, для аналітика в області аналізу даних, саме тут і починається наука. Необхідно переконатися, що в цьому завданні краще працюють евристики, які оцінюють інтенсивність стрибків сигналу, навчитися їх ефективно генерувати (а не перебирати вручну), знайти серед них оптимальну.



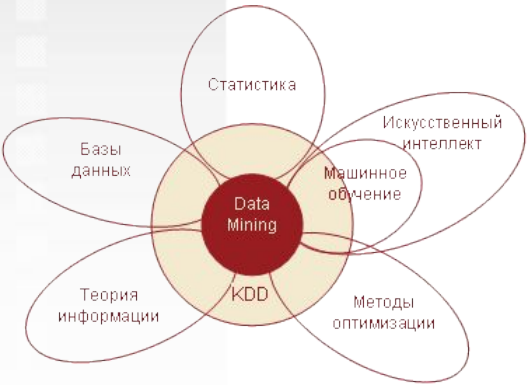
3. Дивовижні закономірності

Не обтяжувати дітей мертвим вантажем фактів, навчіть їх прийомам і способам, які допоможуть їм постягать.
(А.Д.Сент-Екзюпері)

Для того, щоб описаний вище метод не сприймався як «чисте везіння», відзначимо, що весь процес можна автоматизувати, тобто не самим розглядати картинки в придуманих просторах, а довірити це ЕОМ, яка буде генерувати ознаки і оцінювати якість одержуваних признакових просторів за допомогою деякого функціоналу. А ми розберемо, як вручну була вирішена задача класифікації сигналів вже іншої природи.

Постановка

На міжнародному змаганні «Ford Classification Challenge» 2008 [Ford] треба було розробити алгоритм, який розрізняє сигнали датчиків в автомобілі, відповідні нормальній роботі двигуна і несправної роботі.



Кілька сигналів навчальної вибірки змагання зображено на рис. 11.

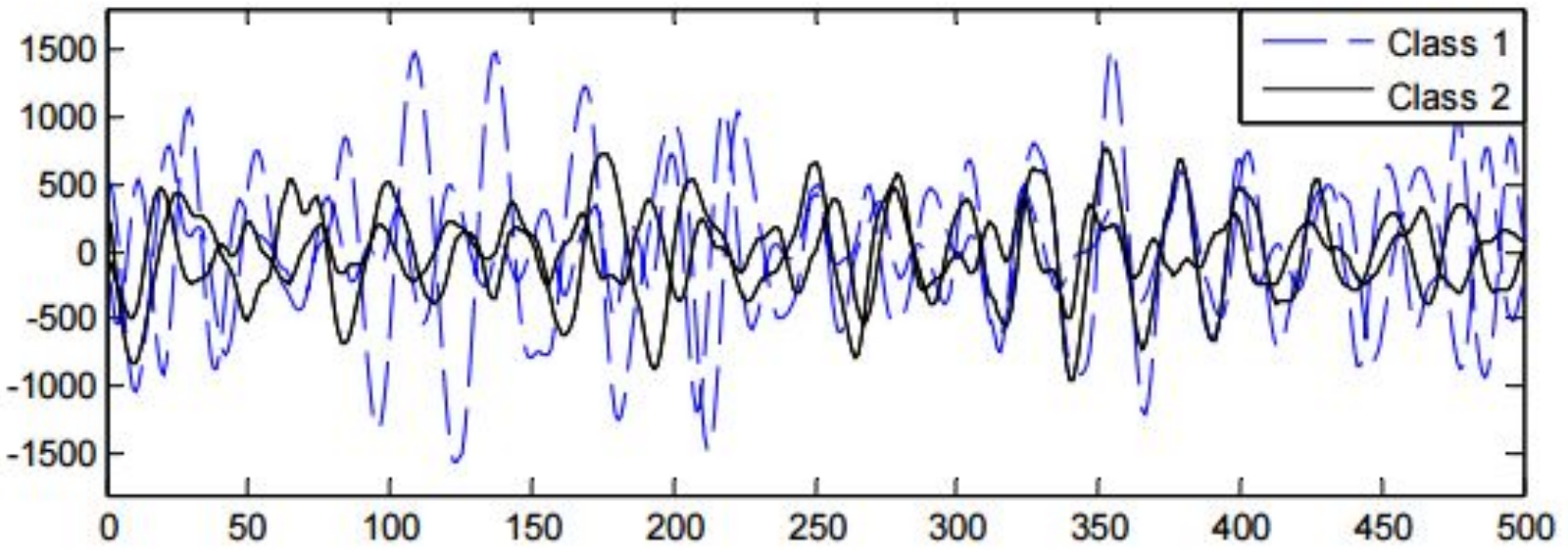
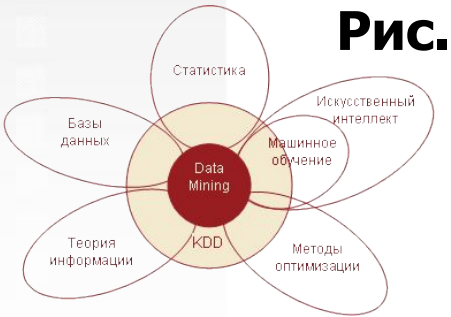


Рис. 11. Сигнали датчиків двигуна (в задачі [Ford]).



У цьому завданні сигнали вже «істотно неоднорідні»: середнє значення другої половини сигналу не залежить від середнього значення першої половини.

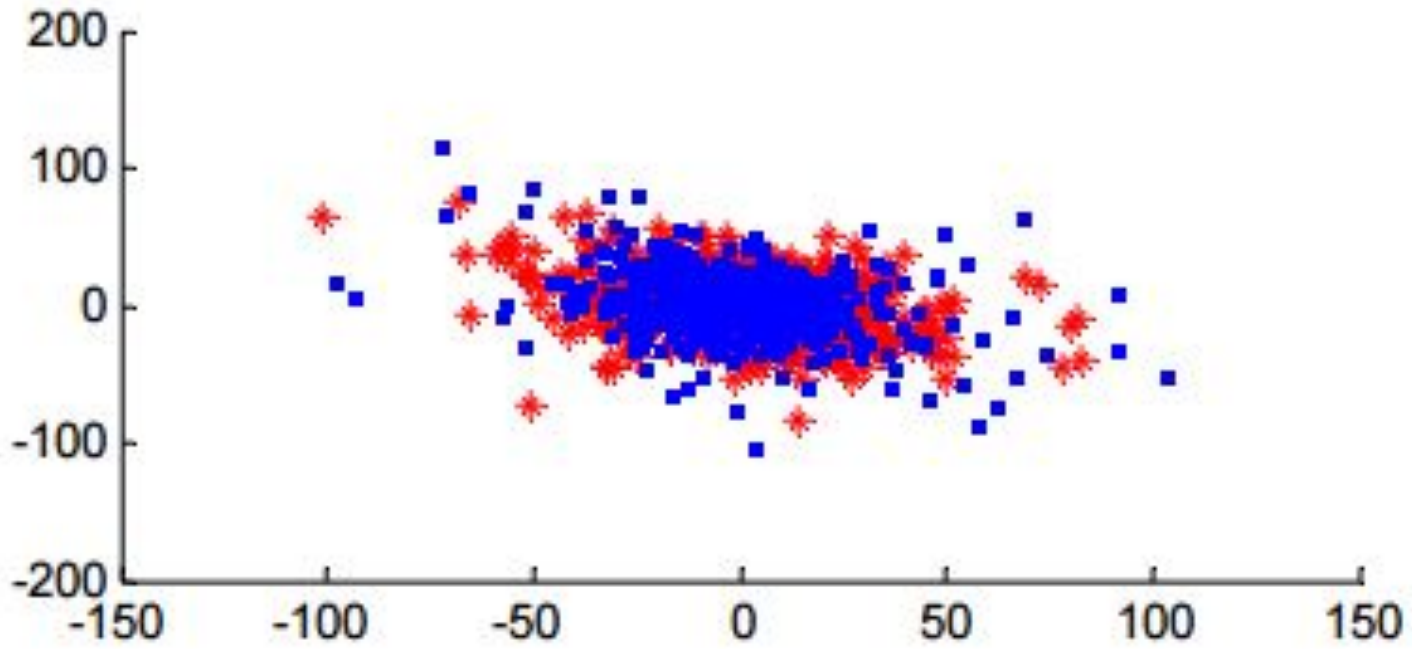
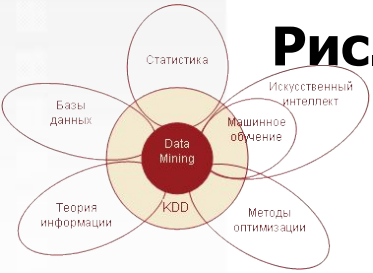
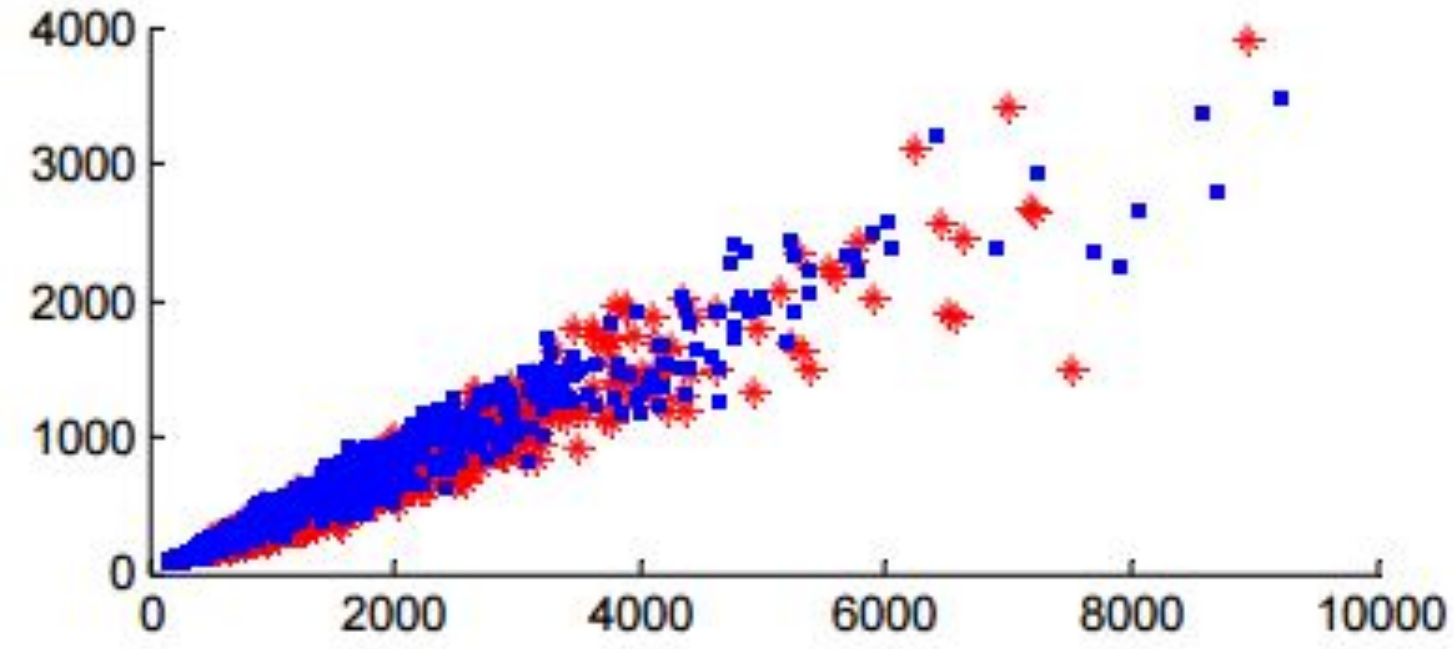


Рис. 12. Середні значення першої половини сигналу (по горизонталі) і останньої (по вертикалі).



У цьому завданні сигнали вже «істотно неоднорідні»: середнє значення другої половини сигналу не залежить від середнього значення першої половини.



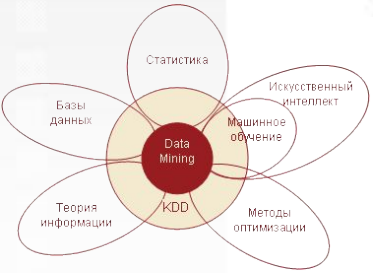


Рис. 13. Максимальні значення сигналів (по горизонталі) і дисперсії (по вертикалі).

Хоча більш явно корелюють максимальні і мінімальні значення (що, до речі, буває досить часто на реальних даних).

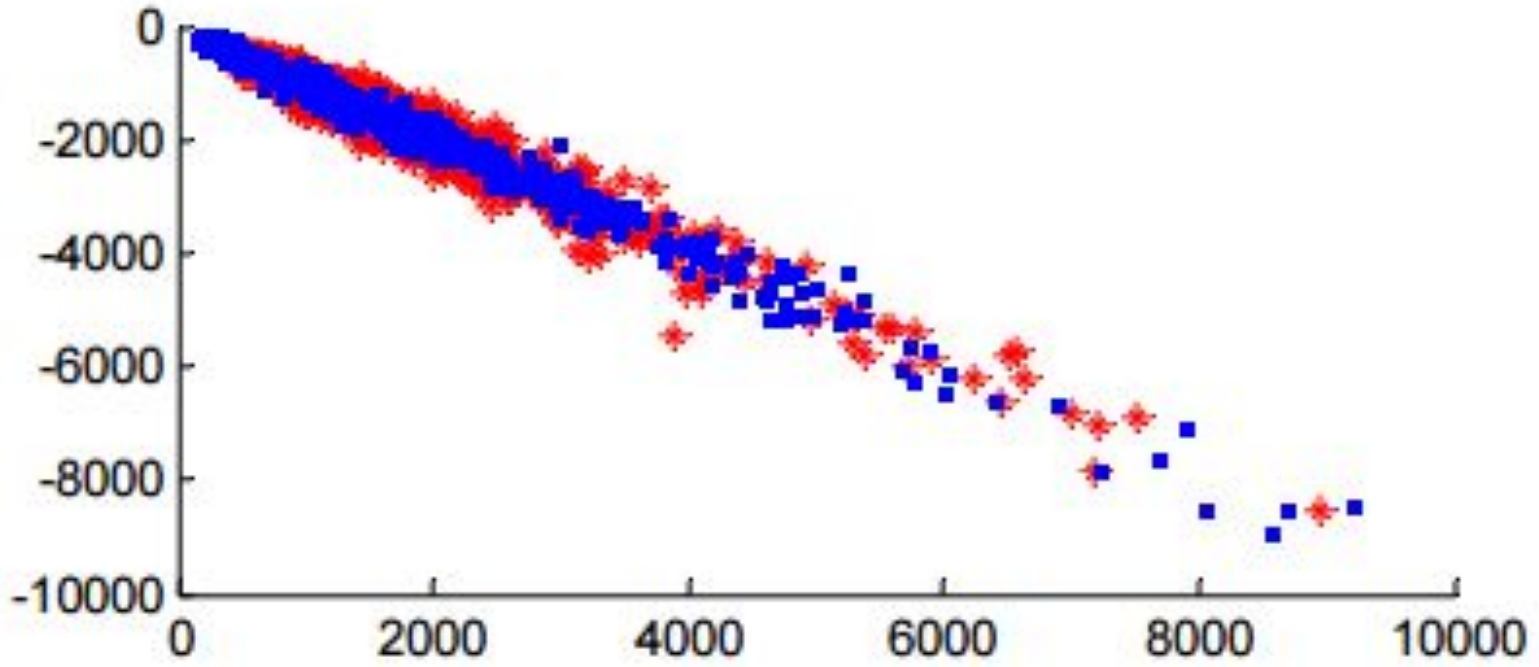
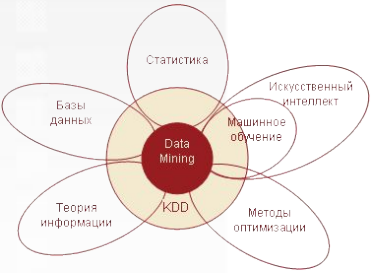


Рис. 14. Максимальні значення сигналів (по горизонталі) і мінімальні (по вертикалі).



Якщо уважно подивитися рис. 14, то видно, що точки одного класу злегка «оточують» точки іншого, а якщо її збільшити, то видно, що частина точок одного з класів утворює щільний згусток. Евристика «якщо максимальне значення сигналу менше 350, то це сигнал першого класу» безпомилково відносить до першого класу 622 сигналу навчання (з 3271).

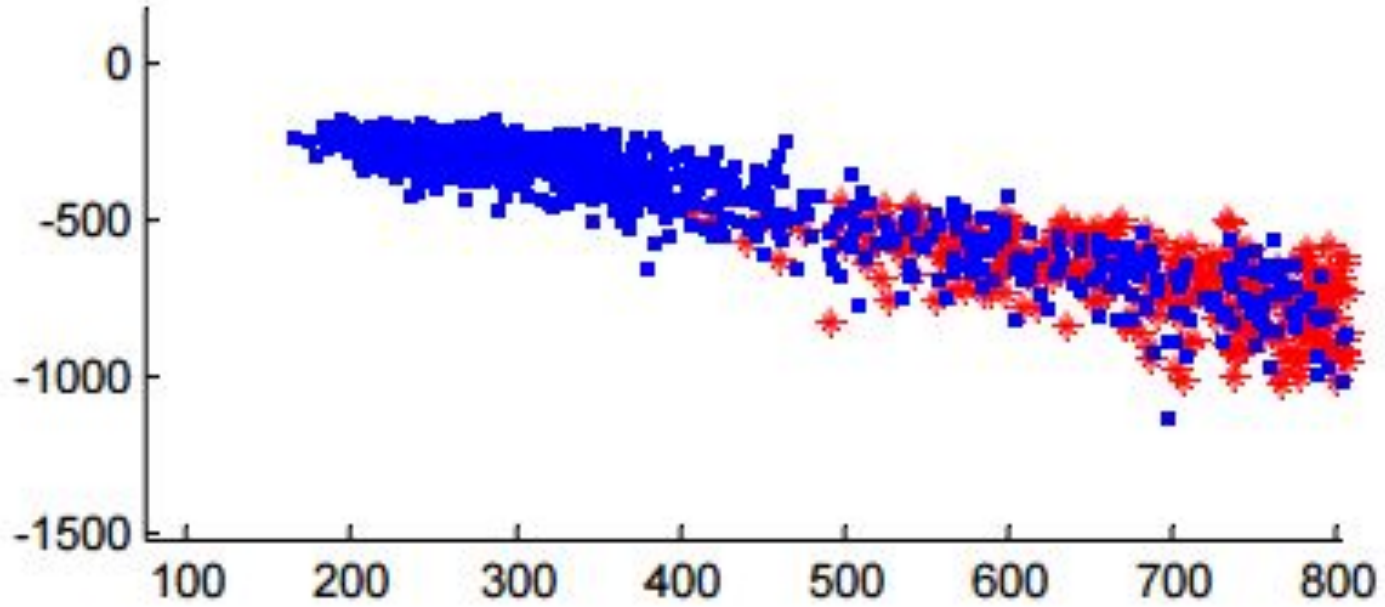
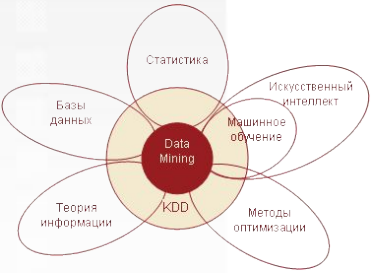
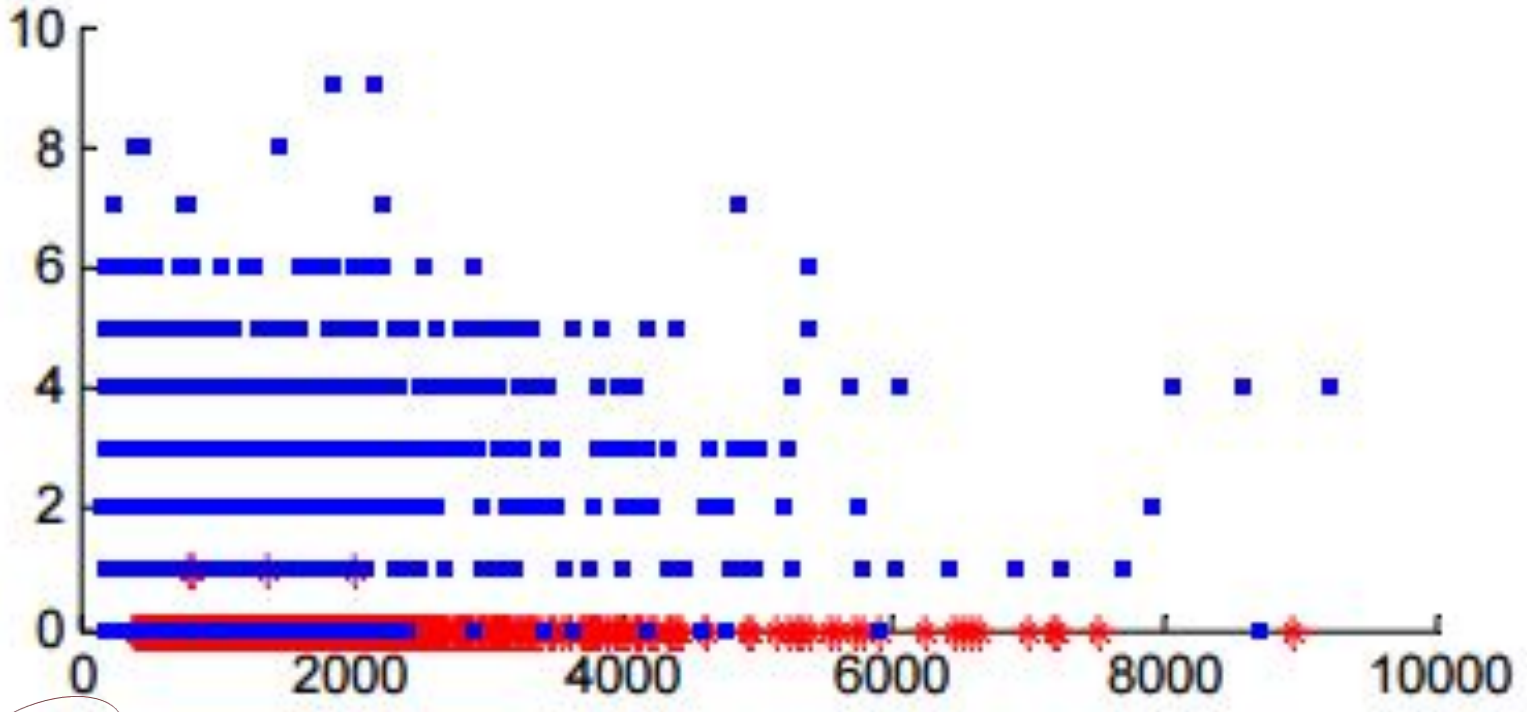


Рис. 15. Збільшений рис. 14.



На рис.16 видно, що також непоганою ознакою виявляється $|\{i \in \{1,2,\ldots,n-1\} \mid u_i = u_{i+1}\}|$ для сигналу $\tilde{u} = (u_1,\ldots,u_n)$,тобто число точок, в яких сусідні значення ui і ui + 1 збігаються. Раз вже ми «намацали» тау непогану ознаку, спробуємо її узагальнити. Перше природне узагальнення – число незначно відрізняється від сусідніх точок. Друге - число повторів значень в сигналі.



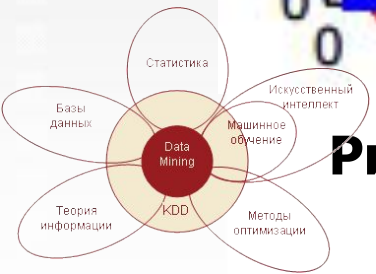


Рис. 16. Максимальні значення сигналів та кількості повторів сусідніх значення в сигналах.

Як видно на рис. 17, друге узагальнення «працює», причому на 100%! Зауважимо, що цей алгоритм реалізується в математичній системі MatLab всього однією командою:

```
2*(sum(diff(sort(X'))==0)<20)'-1).
```

Це і є «шаманство в аналізі даних», коли відповідь завдання криється в 33 символах.

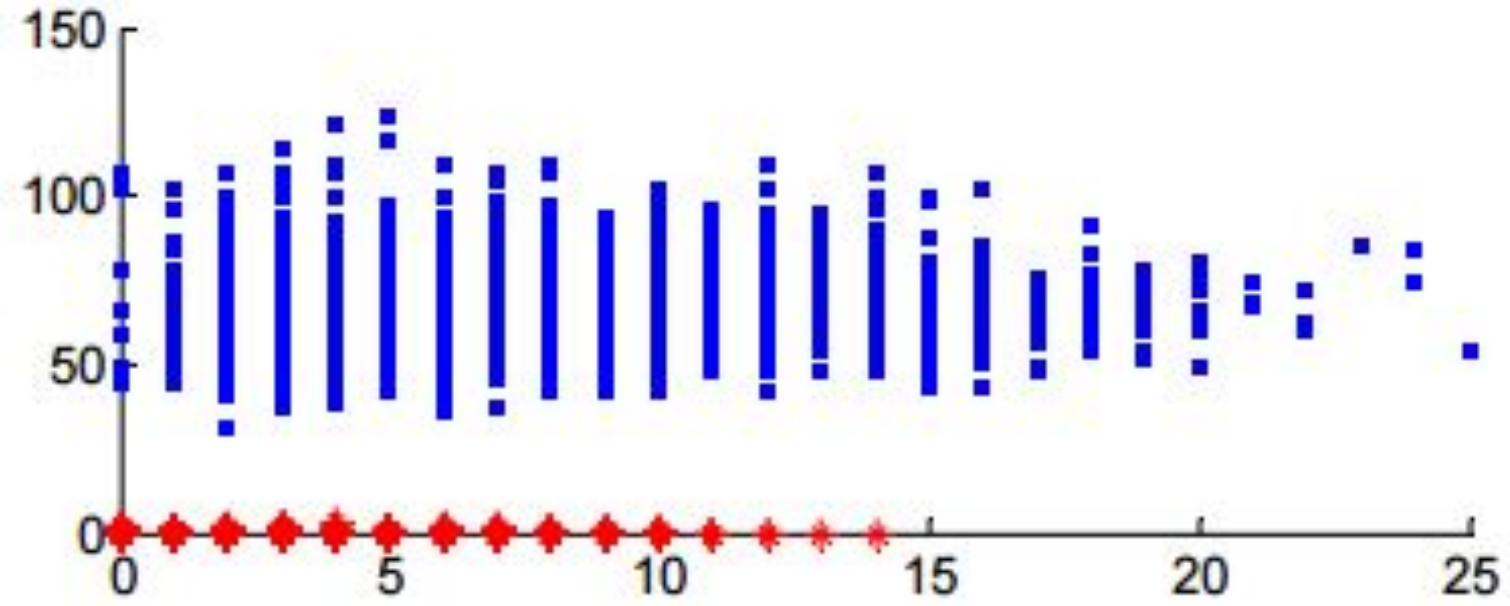
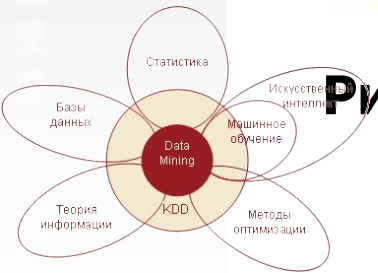


Рис. 17. Перше узагальнення ознаки (по вертикалі) і друге (по горизонталі).



4. Правила справжнього шамана

Дорога до істини вміщено парадоксами
(О. Уайльд)

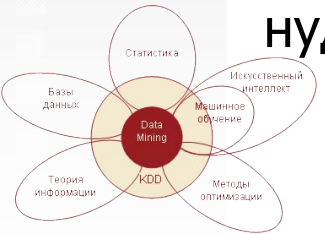
1. Спочатку треба «подивитися на задачу».

До того як застосовувати перевірені (і не дуже) алгоритми треба поглянути, а що з себе представляють реальні дані. Можливо, в них є помилки, видні «неозброєним поглядом». **Тому важливо просто подивитися на дані, зрозуміти, яке значення чому відповідає, зобразити це на графіках.**

Все прекрасне також важко, як і рідко
(Б. Спіноза)

2. У реальної задачі є дуже просте і ефектне рішення.

Скоріше це девіз шамана, якщо він не вірить у наявність «красивого рішення», то завдання вирішувати буде дуже нудно.

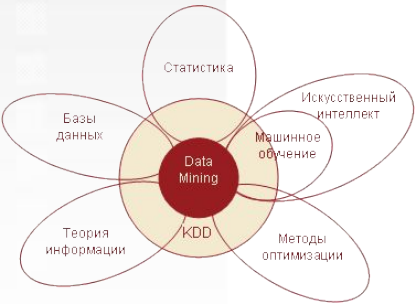


Час не щадить те,
що зроблено без витрати часу.
(Е. Делакруа)

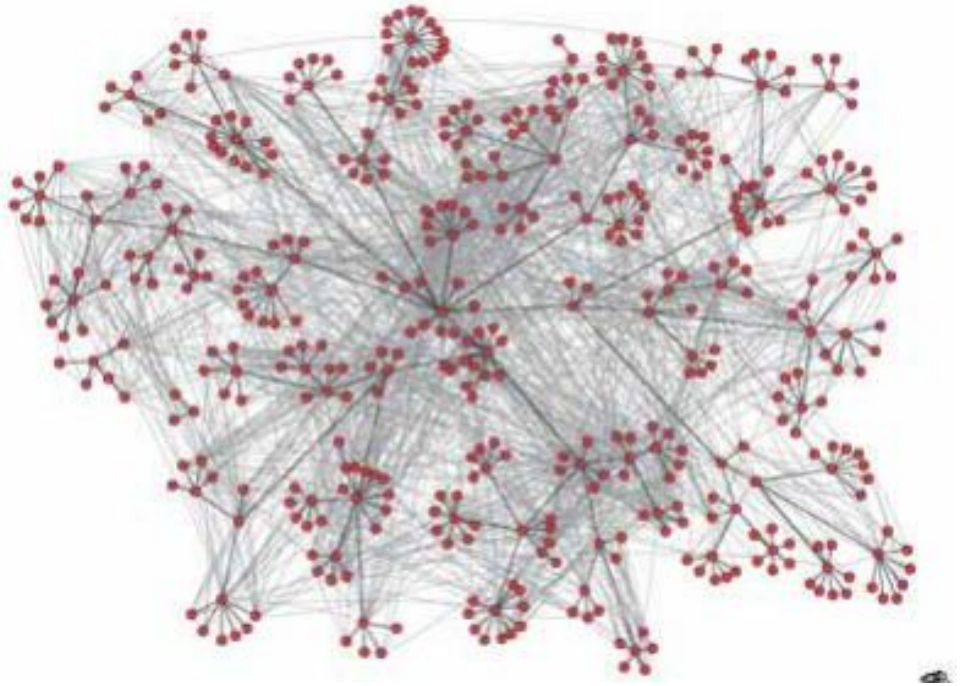
3. Рішення прикладних задач вимагає практики.

Багато в чому, аналіз даних - це дійсно не наука, а ремесло, бо доводиться багато програмувати, причому ефективно програмувати!

Наприклад, в задачі аналізу соціальної мережі доводиться «возитися» з величезним графом (див. Рис. 18). Адже соціальна мережа це граф: користувачі - це вершини, а відносини дружби - ребра. Число вершин може бути більше мільйона, а число ребер - кілька мільйонів. Алгоритм, який аналізує цей граф, не може працювати вічно! Він повинен працювати, як це прийнято говорити, «за прийнятний час».



Ось ще приклад завдання аналізу даних - **прогнозування зв'язності графа.** Необхідно спрогнозувати, які ребра в динамічному (тобто постійно мінливому) графі з'являться найближчим часом. У термінах соціальної мережі - це запропонувати користувачеві «потенційних друзів», тобто людей, з якими швидше за все він знайомий, але ще не «зафрендити».



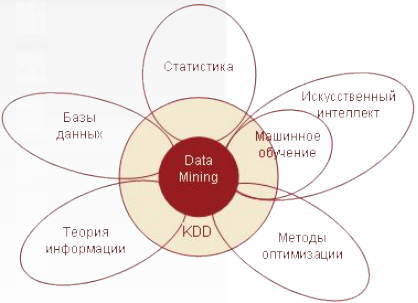


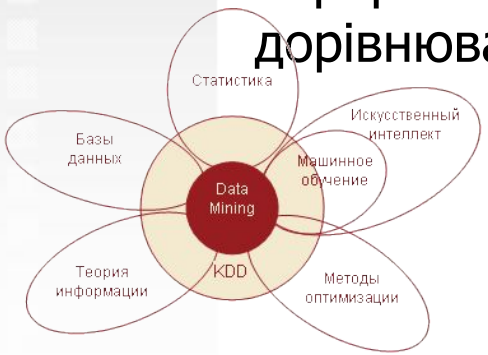
Рис. 18. Граф, відповідний телефонним переговорам

З погляду програмування дуже схожі зовсім різні завдання.

Задача ієрархічної класифікації текстів.

Є новинний ресурс, всі новини якого зберігаються в ієрархічній формі, як файли в операційній системі (див. Рис. 19). Є розділи «спорт», «наука», «політика», «мистецтво» і т.д. в кожному з них є підрозділи (підкаталоги). На ресурс надходять новини, які треба за цими розділами розкладати, тобто повинен бути алгоритм, який аналізує зміст новини і поміщає її в потрібний каталог. Навіть якщо алгоритм буде помилятися, краще щоб він робив це на нижньому рівні ієрархії.

У цьому завданні вихідна інформація представлялася у вигляді гігантської разреженної матриці: по рядках були перераховані тексти, а по стовпцях слова, ij-й елемент дорівнювала кількості входжень j-го слова в i-й текст.



При вирішенні завдання з соціальною мережею, на подив, яка зовсім не схожа на класифікацію текстів, дуже знадобився досвід ієрархічної класифікації. Адже тут теж величезна розріджена матриця: матриця суміжності графа, в ній ij-й елемент дорівнював одиниці, якщо i-й користувач «дружив» з j-м, і нулю в іншому випадку. Причому алгоритми в обох завданнях роблять схожі операції з цими матрицями.

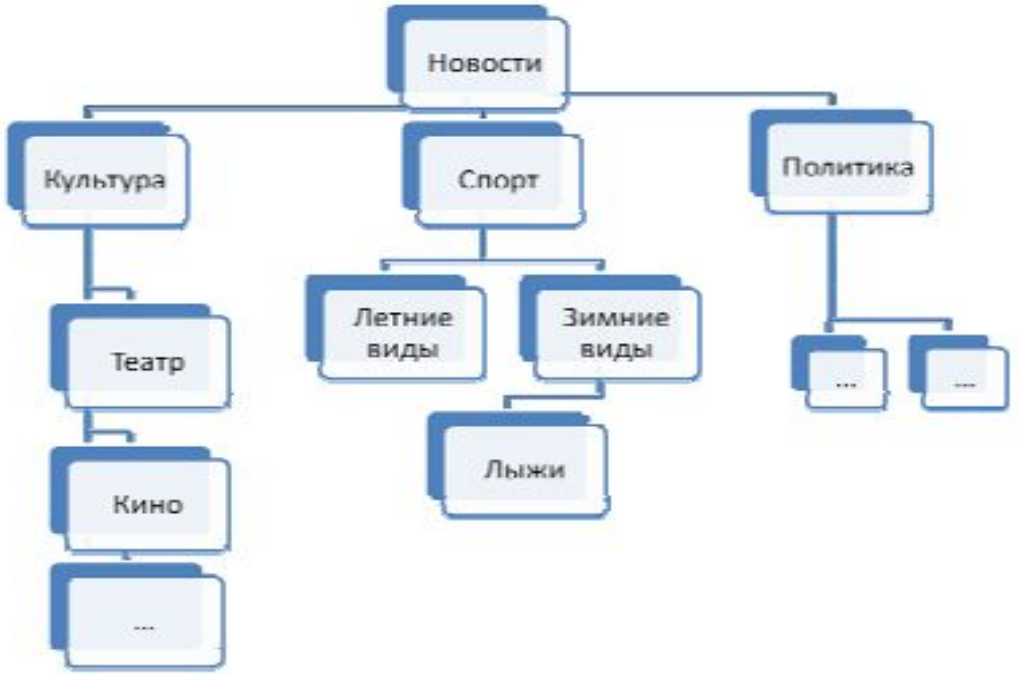


Рис. 19. Каталоги ієрархічної класифікації