Лекція 1. Основні поняття розпізнавання образів

Розпізнавання образів (pattern recognition) — це розділ теорії штучного інтелекту (artificial intelligence), що вивчає методи класифікації об'єктів. За традицією об'єкт, що піддається класифікації, називається *образом* (pattern). Образом може бути цифрова фотографія (розпізнавання зображень), буква або цифра (розпізнавання символів), запис мови (розпізнавання мови) тощо.

В межах теорії штучного інтелекту розпізнавання образів включається в більш широку наукову дисципліну — *теорію машинного навчання* (machine learning), метою якої є розробка методів побудови алгоритмів, що здатні навчатися.

Існує два підходи до навчання: індуктивне і дедуктивне. *Індуктивне навчання*, або *навчання за прецедентами*, засноване на виявленні загальних властивостей об'єктів на підставі неповної інформації, отриманих емпіричним шляхом. *Дедуктивне навчання* передбачає формалізацію знань експертів у вигляді баз знань (експертних систем тощо). В нашому курсі нас буде цікавити лише індуктивне навчання, тому будемо вважати машинне навчання і навчання за прецедентами синонімами.

Слід зауважити, що, як кожна математична дисципліна, розпізнавання образів має власний математичний апарат, який включає математичну статистику, методи оптимізації, дискретну математику, алгебру і геометрію.

Розпізнавання образів має широке застосування і використовується при створенні усіх комп'ютерних систем, на які покладаються інтелектуальні функції, тобто функції, пов'язані із прийняттям рішень замість людини: медична діагностика, криміналістична експертиза, пошук інформації та інтелектуальний аналіз даних тощо.

Прецедент — це об'єкт, приналежність якого до заданого класу визначена заздалегідь. Прецедентом може бути, наприклад, набір ознак пацієнта із відомим діагнозом, з яким слід порівнювати набір ознак людини, діагноз якої ще невідомий.

Кожний образ являє собою набір чисел, що описують його властивості і називаються *ознаками* (feature). Упорядкований набір ознак об'єкта називається *вектором ознак* (feature vector). Вектор ознак — це точка в *просторі ознак* (feature space).

Класифікатор, або вирішальне правило (decision rule) — це функція, яка ставить у відповідність вектору ознак образу клас, до якого він належить.

Задачу розпізнавання образів можна розділити на ряд підзадач.

1. *Генерування ознак* (feature generation) — вимірювання або обчислення числових ознак, що характеризують об'єкт.
2. *Вибір ознак* (feature selection) — визначення найбільш інформативних ознак для класифікації (в цей набір можуть входити не лише первинні ознаки, але й функції від них).
3. *Побудова класифікатора* (classifier construction) — конструювання вирішального правила, на підставі якого здійснюється класифікація.
4. *Оцінка якості класифікації* (classifier estimation) — обчислення показників правильності класифікації (точність, чутливість, специфічність, помилки першого та другого роду).

Введемо позначення і сформулюємо математичну постановку задачі класифікації.

Нехай Ω — простір образів; $\omega \in \Omega$ — образ; $M = \{1, 2, ..., m\}$ — номери класів $\Omega_1, \Omega_2, ..., \Omega_m$, таких що $\Omega_i \cap \Omega_j$, якщо $i \neq j$ і $\bigcup_{i=1}^{m} \Omega_i = \Omega$; $g : \Omega \to M$ — індикаторна функція, що є невідомою; X — простір ознак, тобто векторний простір, точками якого є вектори ознак образів; $x : \Omega \to X$ — функція, що ставить у відповідність образу ω його вектор ознак $x(\omega)$; $K_1, K_2, ..., K_m$ — підмножини простору X, такі що $K_i \cap K_j$, якщо $i \neq j$ і $\bigcup_{i=1}^{m} K_i = X$; $\hat{g} : X \to M$ — вирішальне правило, яке ставить у відповідність вектору ознак образа номер класу, якому він належить.

Задача класифікації з учителем (supervised classification) полягає у тому, щоб на підставі множини прецедентів (g_j, x_j), $j = 1, ..., N$, яка називається *навчальною вибіркою* (training sample) побудувати вирішальне правило $\hat{g}$, що мінімізує кількість помилок. Вчителем вважається або сама навчальна вибірка, або той, хто указав значення g_j.

Задача класифікації без учителя (unsupervised classification) часто називається *кластеризацією* (clusterization), або *кластерним аналізом* (cluster analysis). В цій задачі вибірка образів x_j, $j = 1, 2, ..., N$ розбивається на підмножини, що не перетинаються (кластери), які складаються із схожих один на одного об'єктів, до того ж вимагається, щоб об'єкти із різних кластерів істотно відрізнялися один від одного.

Розглянемо приклад реальної задачі із області медичної діагностики.

У 1962 р. H. Nieburgs, а потім й інші дослідники, повідомили про характерні зміни в клітинах буккального епітелію пацієнта з пухлиною, локалізованої поза порожниною рота, і назвали це змінами, асоційованими з малігнізацією (МАС - malignancy assotiated changes). Ці зміни характеризуються зростанням розмірів ядер, розривами ядерної мембрани, зростанням розмірів зон, пов'язаних із хроматином, оточеним світлими областями. Ці автори зареєстрували схожі зміни при наявності передпухлинних процесів в організмі. У 2009 році пухлинно-асоційовані зміни були індуковані експериментально у мишей, яким привили злоякісні пухлини. Було висунуто гіпотезу, що імунна система реагує на появу злоякісних клітин, виробляючи фактор запалення (inflammation factor), який руйнує ДНК і провокує її згущення. Отже, зареєструвавши такі пухлинно-асоційовані зміни, ми могли б діагностувати рак.

Для дослідження були взяті групи жінок із діагнозами "рак молочної залози" (РМЗ 2-й і 3-й стадії) і "фіброаденоматоз" (ФАМ) у віці від 25 до 53 років (усього 104 пацієнтки). Після полоскання і зняття поверхневого прошарку клітин слизової оболонки порожнини рота були отримані клітини з різноманітної глибини шипуватого прошарку. Ці клітині є *образами*, що піддаються класифікації, а їхня сукупність утворює простір образів, який розбито на два *класи*: РМЗ і ФАМ. Така класифікація називається бінарною.

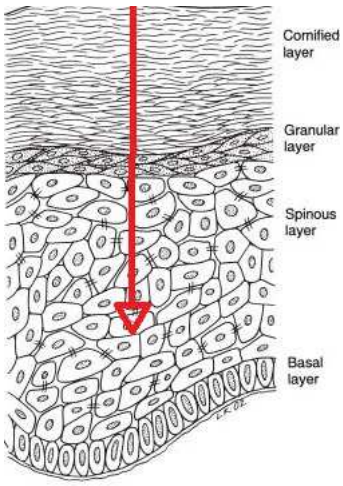


Рис. 1 Структура буккального епітелію (верх — внутрішній бік щоки)

Оптична щільність ядра реєструвалася цитоспектрофотометром за допомогою методу сканування з довжиною хвилі 575 нм і діаметром зонда 0.05 мкм. У кожному препараті досліджувалося від 10 до 30 клітин.

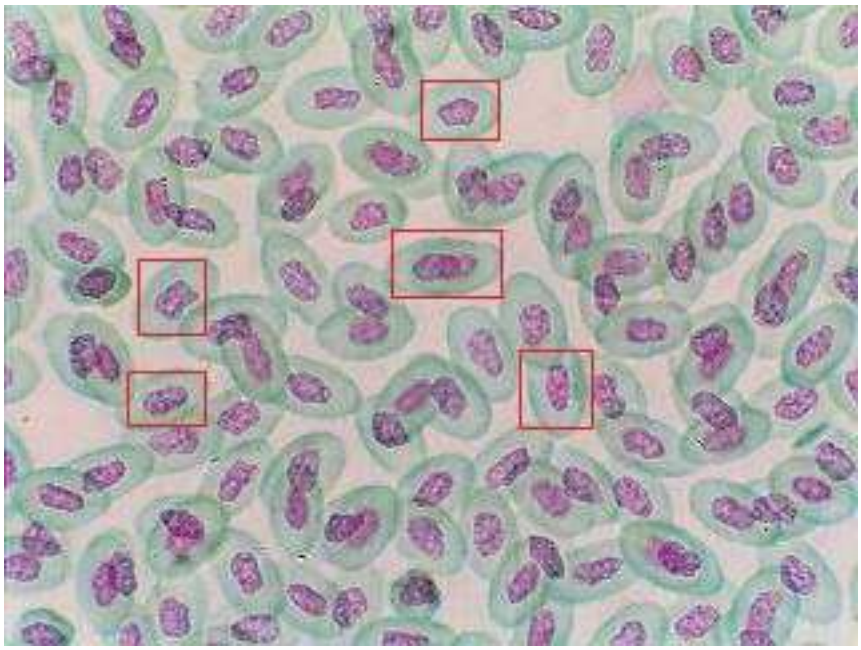
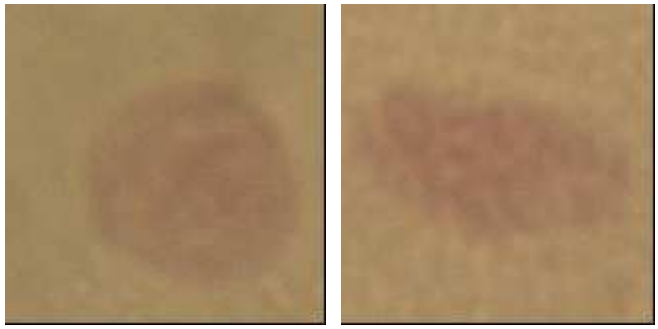


Рис. 2. Ядра клітин після забарвлення за Фьольгеном (рожевим забарвлено ДНК, решта клітин розчиняється і утворює бежевий фон)



а) б)

Рис. 3. Фотографія ядра клітини буккального епітелію після забарвлення за Фьольгеном у жінки, хворої на РМЗ (а) і жінки хворої на ФАМ (б)

При дослідженні інтерфазних ядер були отримані сканограми розподілу ДНК, які є матрицею $R = \left\| r_{ij} \right\|_{i=\overline{1,m}}^{j=\overline{1,n}}$, де r_{ij} характеризує вміст ДНК в осередку сітки з номером (i, j), m — кількість рядків і n — кількість стовпчиків у матриці R. Число $m \times n$ називається розміром реєстраційного поля сканограми.

На підставі цієї цитоспектрофотометричної інформації обчислюються морфо- і денситометричні показники (*ознаки*), що характеризують структурні і текстурні особливості хроматину, наприклад, площа ядра, тобто x_1 — число елементів матриці R, таких що $r_{ij} \geq 0.08$. Кількість таких показників може бути довільною, тому виникає задача *вибору ознак*, яку ми зараз не розглядаємо.

Пропускаючи деталі, які ми розглянемо в подальшому, зауважимо, що задачу було зведено до побудови областей в *просторі ознак*, який мав 15 вимірів, за кількістю показників.

На рис. 4 показано лінійний класифікатор (лінійна дискримінантна функція Фішера) і квадратичний класифікатор (довірчі еліпси).

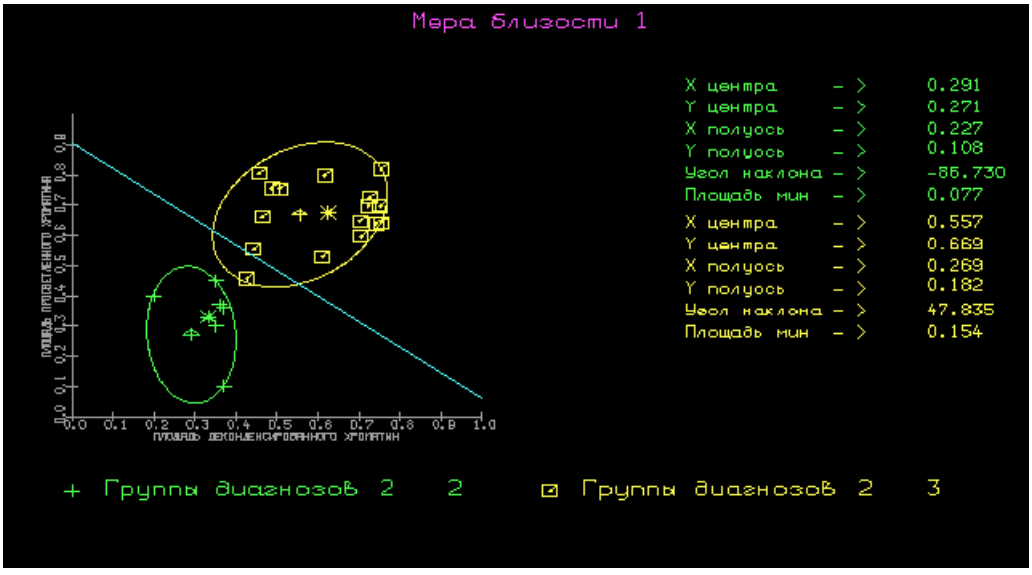


Рис. 4. Лінійний та квадратичний класифікатори

У *вирішальному правилі*, яке було побудовано на підставі лінійних та квадратичних дискримінантний функцій, ураховувалося 210 пар еліпсів та 210 пар напівплощин (це кількість пар показників, що їх можна утворити із 15 показників, не порівнюючи показник сам із собою).

На першій стадії сформуємо дві групи сканограм пацієнтів $A = \{X_i\}_{i=\overline{1,N}}$ і $B = \{Y_j\}_{j=\overline{1,M}}$, діагнози яких точно ідентифіковані. Надалі для певності будемо вважати, що група A містить сканограми хворих із злоякісною пухлиною (рак молочної залози), а група B — із доброякісним новоутворенням (фіброаденоматоз). Після процедур реєстрації і визначення морфо- і денситометричних показників ми одержуємо навчальні вибірки для кожного показника x_k $(k = 1,2, \ldots, 15)$: $G_A^{(1)}, G_A^{(2)}, \ldots, G_A^{(15)}$ для пацієнтів групи A і $G_B^{(1)}, G_B^{(2)}, \ldots, G_B^{(15)}$ для пацієнтів групи B.

Природно припустити, що число вибірок у групах A і B повинно бути однаковим. Для прийняття або відхилення цієї гіпотези ми застосували *процедуру калібрування навчальних вибірок (method one-hold-out),* що складається з таких етапів:

1. Виключимо пацієнта X_i, $i = \overline{1,N}$ (або Y_j, $j = \overline{1,M}$) із множини $A \cup B$.

2. На основі множини вибірок $\{A \cup B\} \setminus X_i$ (або $\{A \cup B\} \setminus Y_j$) побудуємо тести, що використовують пари еліпсів $(E_{ts}, \overline{E}_{ts})$, $(E_{ts}^*, \overline{E}_{ts}^*)$ і напівплощин (π_{ts}, λ_{ts}), $(\pi_{ts}^*, \lambda_{ts}^*)$.

3. Обчислимо значення статистик $h_k = h(C_k)$ $(k = \overline{1,6})$ для пацієнта X_i, $i = \overline{1,N}$ (або Y_j, $j = \overline{1,M}$), де h_k — це частота певної події (наприклад, потрапляння в певну частину еліпса або у певну напівплощину). Деталі зараз не важливі.

4. Повернемо пацієнта X_i, $i = \overline{1,N}$ (або Y_j, $j = \overline{1,M}$) у множину $A \cup B$ і повторимо процедуру для іншого пацієнта.

Введемо в розгляд такі критерії діагностики:

1) квадратичний: $h_3 > h_4 \Rightarrow$ РМЖ; $h_3 \le h_4 \Rightarrow$ ФАМ;

2) лінійний: $h_5 > h_6 \Rightarrow$ РМЖ; $h_5 \le h_6 \Rightarrow$ ФАМ.

Це — *вірішальне правило*. Як бачимо, воно може бути дуже складним.

Введемо такі позначення: D_1 - діагноз "РМЖ", D_2 - діагноз "ФАМ", ν_{11} - частота події D_1 серед вибірок РМЖ, ν_{21} - частота події D_2 серед вибірок РМЖ, ν_{12} - частота події D_1 серед вибірок ФАМ, ν_{22} - частота події D_2 серед вибірок ФАМ.

Аналіз результатів, отриманих при калібруванні вибірок із груп A і B однакового обсягу (табл.1) дозволяє зробити такі висновки.

1. У переважній більшості випадків для групи В (навчальна вибірка пацієнтів із ФАМ) спостерігається перевищення (домінування) статистики h_4 (тотальний ФАМ) над h_3 (тотальний РМЖ), а також статистики h_2 (ФАМ)

над h_1 (РМЖ) (ми будемо називати цей феномен *ефектом стійкого домінування*), а для групи A такого ефекту не спостерігається.

Для лінійного критерію події D_1 і D_2 практично рівноймовірні, як для групи A (навчальна вибірка пацієнтів із РМЖ), так і для групи B (навчальна вибірка пацієнтів із ФАМ), тому цей критерій непридатний для діагностики РМЖ і ФАМ.

Таблиця 1. Частота випадкової події D_k $(k = 1,2)$ при калібруванні навчальних вибірок (25 РМЗ і 25 ФАМ)

Частоти Критерії	ν_{11}	ν_{21}	ν_{22}	ν_{12}
Квадратичний	0.28	0.72	0.80	0.20
Лінійний	0.56	0.44	0.48	0.52
Комбінований	0.72	0.28	0.80	0.20

Ця таблиця демонструє якість розпізнавання.

Чутливість (ν_{11}) — це доля правильно розпізнаних ракових хворих. Вона обчислюється за формулою: (кількість правильно розпізнаних хворих на рак)/(загальна кількість хворих на рак).

Специфічність (ν_{22}) — це доля правильно розпізнаних хворих на альтернативну хворобу, у даному випадку це ФАМ. Вона обчислюється за формулою: (кількість правильно розпізнаних хворих на ФАМ)/(загальна кількість хворих на ФАМ).

Точність — це доля правильно розпізнаних об’єктів незалежно від діагнозу, тобто (кількість правильно розпізнаних хворих на рак + кількість правильно розпізнаних хворих на ФАМ)/(загальна кількість пацієнтів). В даному випадку, оскільки кількість хворих на рак і ФАМ є однаковим, це число дорівнює напівсумі чутливості і специфічності : $(\nu_{11}+\nu_{22})/2$. Але якщо кількість об’єктів не однакова, слід застосовувати загальну формулу.

Ось так на практиці застосовуються основні поняття розпізнавання образів.

Контрольні питання

1. Що таке розпізнавання образів?
2. Що таке машинне навчання?
3. Що таке образ? Наведіть приклади образів.
4. Назвіть різновиди навчання.
5. Що таке прецедент?
6. Що таке ознака, вектор ознак і простір ознак?
7. Що таке вирішальне правило?
8. На які підзадачі розбивається задача класифікації?
9. Назвіть різновиди класифікації.
10. Сформулюйте математичну постановку задачі класифікації з учителем.
11. Опишіть приклад класифікації, викладений в лекції, використовуючи терміни математичної постановки задачі: простір образів, простір ознак тощо.
12. Що таке чутливість, специфічність і точність бінарної класифікації?

Перелік джерел

1. Дуда Р., Харт П. Распознавание образов и анализ сцен. — М.: Мир, 1976.
2. Фукунага К. Введение в статистическую теорию распознавания образов. — М.: Наука, 1979.
3. Главач В., Шлезингер М.И. Десять лекций по статистическому и структурному распознаванию образов. К.: Наукова думка, 2004. www.irtc.org.ua/image/Files/Schles/esh10_full.pdf.
4. Воронцов К.В. Машинное обучение. (Курс лекций). ВмиК МГУ: Москва, 2009. http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций%2C_К.В.Воронцов)
5. Местецкий Л.М. Математические методы распознавания образов. (Курс лекций). ВмиК МГУ: Москва, 2004). www.ccas.ru/frc/papers/mestetskii04course.pdf.
6. Лепский А.Е., Броневич А.Г. Математические методы распознавания образов. (Курс лекций). Южный федеральный университет: Таганрог, 2009. http://www.lepskiy.ucoz.com/lect_Lepskiy_Bronevich_pass.pdf
7. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов М.: Наука, 1974. — 416 с.