# Побудова та дослідження моделі навчання метричних класифікаторів на основі ймовірнісно-комбінаторного підходу

Борис Капустій, Богдан Русин, Віталій Таянов

Відділ методів та систем аналізу обробки та ідентифікації зображень
Фізико-механічний інститут ім. Г.В. Карпенка НАН України
vtayanov@ipm.lviv.ua

В роботі представлена повна завершена концепція підходу, що дозволяє оцінити вплив навчаючої вибірки на результати розпізнавання при використанні метричних класифікаторів. Таким чином вдається визначити, наскільки розмір навчаючої вибірки впливає на параметри моделі алгоритмів класифікації. Отримані результати будуть використані для більш точної та адекватної оцінки кількості надлишкової інформації, що міститься у навчаючих даних, що дозволить зменшити ефект перенавчання. У випадку систем розпізнавання, що використовують навчання розмір навчаючої вибірки та її склад є основними параметрами самого процесу навчання, а отже, вони визначають основні технічні характеристики системи.

## 1. Вступ

Як відомо, всі класифікуючі алгоритми можуть бути поділені на три групи: алгоритми із навчанням, із самонавчанням та алгоритми, що не використовують навчання як такого. Найбільш важливими серед цієї групи є алгоритми, що використовують навчання. Ці алгоритми є об'єктом дослідження в рамках теорії машинного навчання (Theory of Machine Learning), яка доволі швидко і успішно розвивається на протязі останніх десяти років [4]. В рамках цієї теорії розглядаються такі важливі питання, як визначення оптимального складу навчаючої вибірки, питання навчання класифікаторів та побудови оптимальної композиції класифікаторів, що задовольняє певним умовам, а також генерації та селекції найбільш інформативних ознак. Алгоритми, що дозволяють певною мірою вирішувати ці питання носять назви Bagging, Boosting та Random Space Method (RSM). Аналіз цих алгоритмів встановлює одну спільну їх рису, спрямовану на зменшення надлишковості та неінформативності як у самих даних (визначення найбільш оптимального складу навчаючої вибірки та набору найбільш інформативних ознак), так і надлишковості (складності) самого апарату класифікації, тобто, власне, самих класифікуючих алгоритмів. Тому потрібно спочатку визначити вплив навчаючих даних на процес розпізнавання з тим, щоб потім провести генерування та селекцію найбільш інформативних ознак та налаштування параметрів класифікатора таким чином, щоб мінімізувати перенавчання і досягти найбільшого значення ймовірності правильного розпізнавання.

## 2. Підходи щодо оцінки якості роботи класифікаторів

Загалом якість роботи класифікаторів прийнято характеризувати через поняття відступу (margin). Відступ визначається як відстань об'єкта від розділювальної гіперплощини. Чим більший відступ, що забезпечується тим або іншим класифікатором, тим кращою вважається робота цього класифікатора. Однак, якщо всі об'єкти або переважна їх більшість мають приблизно однаково великий відступ і групуються один біля одного, то в цьому випадку різко падає їх інформативність. Це означає, що замість всіх об'єктів можна залишити один або декілька, що використовуються для навчання. На цьому прикладі показано одну з основних причин, що обумовлюють ефект перенавчання. Однобічне налаштування алгоритму на основі близької за суттю навчаючої інформації призводить до того, що в реальних умовах алгоритм буде максимально часто помилятись, навіть якщо він не мав помилок на навчаючій вибірці. Дійсно, ймовірність того, що нам зустрінеться така сама ситуація, яка була в умовах навчаючої вибірки, є близькою до нуля. Тому прийнято використовувати для навчання несхожі і "важкі"для алгоритму об'єкти з малими значеннями відступу. Ця ідея використана, наприклад, у методі опорних векторів (Support Vector Machine) або методі зваженого голосування. Також використовують узагальнений підхід для характери-

стики класифікаторів на основі поняття відступу. Результатом роботи метричних класифікаторів є ранжовані дані (посортовані за ступенем подібності до тестового об'єкта об'єкти бази даних). При цьому для таких класифікаторів поняття відступу представляється по-іншому. Вводиться еквівалентна до відступу характеристика, яка для даного об'єкта може бути представлена, як відносна відстань між його відстанню від тестового об'єкта та аналогічною відстанню від усередненого об'єкта бази даних або до останнього об'єкта з однорідної (стратегічної)[1] послідовності своїх об'єктів. Передбачається, що хоча б частина своїх об'єктів розташовуються на початку списку можливих претендентів. Таким чином, гарантується коректність даного означення.

### 2.1. Аналіз процесу класифікації при використанні метричних класифікаторів

Під метричним класифікатором розуміють відображення виду

$$a(u; X^\ell) = \arg\max_{y \in Y} \underbrace{\sum_{i=1}^{\ell} [y_{i,u} = y] w(i,u)}_{\Gamma_y(u, X^\ell)}.$$

Для такого класифікатора його робота обумовлюється тим, що рішення про клас приймається на основі максимальної сумарної ваги $\Gamma_y(u) \equiv \Gamma_y(u, X^\ell)$. Ще одною перевагою метричних класифікаторів, крім їх простоти, на нашу думку, є те, що рішення, прийняте цими класифікаторами не залежить від порогу. Сама задача вибору порогу є трудомісткою, складною і вимагає численних тренувань на основі достатньо великої навчаючої вибірки. Разом з тим метричні класифікатори мають достатню кількість ступенів свободи для їх налаштування і вони є, як правило, більш стійкими до впливу зовнішніх факторів за рахунок їх інтегрального характеру, ніж порогові класифікатори.

Серед метричних класифікаторів за ступенем збільшення складності можна відзначити наступні

- $w(i,u) = [i = 1]$ – алгоритм найближчого сусіда;
- $w(i,u) = [i \le k]$ – алгоритм $k$ найближчих сусідів;
- $w(i,u) = [i \le k] q^i$ – зважений алгоритм $k$ найближчих сусідів;
- $w(i,u) = K\left(\frac{\rho(u,x_{i,u})}{h}\right)$ – парзеновське вікно фіксованої ширини;
- $w(i,u) = K\left(\frac{\rho(u,x_{i,u})}{\rho(u,x_{k+1,u})}\right)$ – парзеновське вікно змінної ширини;
- метод потенційних функцій.

У випадку алгоритму найближчого сусіда $k = 1$. Для алгоритму $k$ найближчих сусідів ваги рівні 1. Для випадку зваженого алгоритму $k$ найближчих сусідів, чим далі об'єкт знаходиться від початку списку можливих претендентів, тим менша його вага. Постає питання про відношення між вагами двох сусідніх об'єктів ($\frac{w_i}{w_{i+1}}$) в списку можливих претендентів. Покажемо, що воно повинно бути в межах $1 \le \frac{w_i}{w_{i+1}} \le 2$. При $\frac{w_i}{w_{i+1}} = 1$ маємо звичайний $k$NN алгоритм, при $1 < \frac{w_i}{w_{i+1}} < 2$ – зважений $k$NN, а при $\frac{w_i}{w_{i+1}} \ge 2$ – алгоритм найближчого сусіда або 1NN. Якщо вага об'єкта пропорційна до ймовірності його незаміщення в списку можливих претендентів об'єктами інших класів, то відбувається поєднання рангового голосування та методу Парзена, що в решті решт і предсталяється як віконний метод Парзена. Основна ідея методу вікна Парзена, полягає в тому, що вага об'єкта задається не його рангом, а на основі функції відстані. При цьому вага об'єкта обчислюється з використанням ядерних функцій з постійним або змінним вікном $h_i$, а центр ядра знаходиться в самому класифікованому об'єкті. Оскільки у методі Парзена ваги на об'єкти визначаються не рангом а відстанями класифікованого об'єкта від навчаючих, то відносна відстань оцінена за параметром $\hat{z}$ та функція розподілу ймовірностей $P(\hat{z})$, між якими є однозначна відповідність, повністю визначають даний алгоритм. Якщо використовувати в якості ядра радіальну базисну функцію, то дисперсія нормального розподілу відстаней відіграє роль вікна $h_i$ у класичному методі Парзена. Перевага такого підходу порівняно з класичним полягає в тому, що розмір вікна автоматично задається складом навчаючої вибірки і "зашитий"в параметрі $\hat{z}$, а також функції $P(\hat{z})$. Метод потенційних функцій є модифікацією алгоритму Парзена, основна відмінність якого полягає в тому, що центр ядра знаходиться не в класифікованому об'єкті, а в навчаючих, тобто використовується набір ядер з різними розмірами вікон $h_i$. Передбачається, що ядра в обох методах є фінітні, оскільки в протилежному випадку для класифікації об'єкта доведеться використовувати всю навчаючу вибірку. Оскільки розподіл відстаней між класифікованим об'єктом і навчаючими є нормальним (або в силу оптимальності повинен таким бути [3]), то обидва підходи (класичний і пропонований) є абсолютно еквівалентні відносно задачі класифікації. Принципова відмінність полягає в тому, що параметри ядер в пропонованому підході визна-

чаються на основі навчаючої вибірки, а також є функціями процесу відбору ознак, способом обчислення відстаней тощо.

## 3. Формалізація задачі

Нехай $X$ – простір об'єктів; $Y$ – множина імен класів, $y^* : X \to Y$ – цільова функція, значення якої відомі лише на об'єктах скінченої навчаючої вибірки $X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$, $y_i = y^*(x_i)$ [5]. У базі даних існують класи еталонів $C_i$, $i = \overline{1,n}$, причому $s_i = |C_i|$ – розміри класів. Передбачається, що розміри $s_i$ всіх класів однакові і рівні $s$. Оскільки існує вибірка контрольних образів $U$, що подаються на розпізнавання, то загальна кількість образів, що приймають участь у процесі розпізнавання, дорівнюватиме $ns + |U|$. Нехай оцінена частота помилок алгоритму класифікації $a = \mu(X^\ell)$ на навчаючій вибірці $X^\ell \subseteq X^L$: $\nu(a, U) = \frac{1}{|U|}\sum_{x\in U}[a(U) \neq y^*(U)]$. Задача полягає в оцінюванні величини

$$\tilde{\nu}(\tilde{a}, U) = \frac{1}{|U|}\sum_{x\in U}[\tilde{a}(U) \neq y^*(U)] \qquad (1)$$

при пониженні розмірності $C_i$ класів-еталонів, де $\tilde{a} = \mu(X^{\tilde{\ell}})$ – алгоритм, побудований на основі вибірки розміру $\tilde{\ell}$.

В якості алгоритму класифікації використаємо алгоритм $k$NN. При такій загальній постановці задачі найбільш придатним підходом до її вирішення є комбінаторний підхід. Очевидно, що в кожному конкретному випадку пониження навчаючої вибірки на основі класів-еталонів буде проводитись не обов'язково оптимальним чином, однак загальна статистика всіх можливих понижень класів та результатів таких понижень має дати відповідь на питання про зменшення інформаційного опису класів на основі еталонних об'єктів у цілому.

## 4. Суть імовірнісно-комбінаторного підходу

Основна суть поєднання двох підходів полягає в тому, щоб досягнути більшої точності і коректності у побудові оцінок ймовірності розпізнавання при зменшенні розміру навчаючих даних. Оцінки ймовірності правильного розпізнавання для малих вибірок розглянуті в [2].

Представимо результати розпізнавання у вигляді двійкової послідовності посортованих за мінімумом відстані об'єктів, де 1 ставиться у відповідність образам, які підтримують успішне розпізнавання, а 0 - образам, які заважають успішному розпізнаванню. Приклад такої послідовності поданий на рис.1. Розглянемо випадок$ENT(\frac{k}{2})+1 \leq s^*$. Визначимо ймовірності того, що серед послідовності образів свого класу даної довжини будуть вибрані комбінаторним способом $s^*$ образів. Такі ймовірності носять довірчий характер і характеризують ступінь накриття нестиснутого класу послідовністю з $|\bigcup_i \ell_i|$ образів, серед яких вибирається $s^*$. Крім них, знайдемо також імовірності того, що не будуть вибрані відповідним способом певні образи з чужих класів. Ймовірність коректної роботи $k$NN класифікатора є добутком цих імовірностей. Визначимо ймовірність помилкової класифікації, обумовленої образами з чужих класів:

$$q_j = \frac{1}{C_s^{s^*}} \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_{i,j} m_{i+j-1}|} C^{s^*-j}_{s-|\bigcup_{i,j} m_{i+j-1}|}; \qquad (2)$$

$$|\bigcup_{i,j} m_{i+j-1}| > ENT(\frac{k}{2}) + 1.$$

Обчислимо довірчу ймовірність для довільної послідовності з образів свого класу:

$$P_{q_j} = \frac{1}{C_s^{s^*}} \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i \ell_i|} C^{s^*-j}_{s-|\bigcup_i \ell_i|}. \qquad (3)$$

Ймовірність правильного розпізнавання при застосуванні $k$NN класифікатора визначається добутком імовірності (3) та доповнення до ймовірності (2):

$$P_i = P(q_i)(1 - q_i) =$$
$$= \frac{1}{C_s^{s^*}} \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i \ell_i|} C^{s^*-j}_{s-|\bigcup_i \ell_i|} - \qquad (4)$$
$$- \frac{1}{(C_s^{s^*})^2} \Big( \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i \ell_i|} C^{s^*-j}_{s-|\bigcup_i \ell_i|} \Big)$$
$$\Big( \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i m_i|} C^{s^*-j}_{s-|\bigcup_i m_i|} \Big).$$

Роль імовірнісної частини в комбнаторно-ймовірнісному підході полягає у тому, що необхідно обчислити ймовірність існування однорідних послідовностей виду $\{0\}$ або $\{1\}$.Обчислення ймовірності існування послідовностей змішаного типу не має сенсу, оскільки для великих розмірів послідовностей вона оберненопропорційна до величини $2^{|\ell+m|}$, де $|\ell + m|$ – розмір послідовності. Ймовірність існування однорідної послідовності з

образів свого класу $\{1\}$ обчислюється на основі ймовірності заміщення останнього образа свого класу у цій послідовності. Тобто розмір однорідної послідовності вказаного виду визначається найбільш "слабким"образом. Отже, потрібно обчислити ймовірність існування заданого розміру послідовності образів свого класу або для заданого рівня ймовірності обчислити максимальний розмір послідовності, який забезпечить цю ймовірність. Для двійкової послідовності сума ваг молодших розрядів завжди на 1 менша за наступний старший розряд. Тобто заміщення довільного образа свого класу у списку еквівалентне почерговому заміщенню усіх попередніх. Мінімальний цілий порядок системи числення, що володіє цією властивістю, рівний 2. Отже потрібно обчислити ваги положень образів свого класу і порівняти їх з двійковими розрядами. Таке представлення дозволяє спростити обчислення ймовірності заміщення образів послідовностей зі своїх класів образами з чужих класів. З іншого боку, довільні ваги можна виразити через показник степені 2, що також спрощує представлення і обчислення цих імовірностей. Таким чином, ймовірність існування однорідної послідовності з образів свого класу обчислюється на основі розподілу відстаней і є функцією від параметрів алгоритму. Приймається така послідовність, для якої ймовірність її існування є достатньою. А далі застосовується комбінаторний підхід, який дозволяє обчислити ефект впливу пониження розміру класів на ймовірність розпізнавання. Оскільки ймовірнісна частина даного підходу визначається параметрами алгоритму розпізнавання, то поєднання ймовірнісної та комбінаторної частини дозволяє більш точно описати ефект від пониження кількості навчаючих даних. Наприкінець покроково розглянемо приклад швидкого обчислення ймовірності заміщення свого образу з послідовності, де співвідношення між вагами об'єктів є ціла степінь числа 2. Отже, наприклад, ваги задаються наступним чином: $w = \{2^9, 2^6, 2^4, 2^3, 2^2, 2^1, 2^0\}$. Як відомо, ймовірність заміщення свого об'єкта з послідовності чужим об'єктом, коли відомо, що заміщення відбулося, оберненопропорційна до ваг цих об'єктів. Знайдемо ймовірність заміщення об'єкта з вагою $2^9$ порівняно з вагою $2^6$. Оскільки невідомо, заміщення якого об'єкта відбулося, то сумарна вага того, що це не будуть об'єкти з вагою $2^6$ і нижче дорівнюватиме: $2^6 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0$. У долях ваги $2^6$ це з точністю до 1 рівне $2^6 * (1 + 0.5) = 1.5 * 2^6$. У випадку великих послідовностей ця 1 мало впливає на точність. Співвідношення між $2^9$ та $2^6$ рівне 8. У випадку повної групи подій отримаємо $8\lambda + 1.5\lambda = 1$, звідки коефіцієнт пропорційності $\lambda$ приблизно рівний 0.11. Таким чином, ймовірність незаміщення об'єкта з вагою $2^9$ рівна $8 * 0.11 = 0.88$, а об'єкта з вагою $2^6$ відповідно $1 - 0.88 = 0.12$. Оскільки у нашому випадку точно відомо, що заміщення відбулося, а останній об'єкт має вагу 1, то поправка на точність, рівна 1, вносить потрібну корекцію.

$$\underbrace{\underbrace{1111}_{l_1}\underbrace{000}_{m_1}\underbrace{111}_{l_2}\underbrace{00}_{m_2}\underbrace{1111}_{l_3}\underbrace{000}_{m_3}\underbrace{111}_{l_4}\ldots\underbrace{000}_{l_n}\ldots\underbrace{111}_{m_n}\ldots}_{\{l,m\}}$$

Рис. 1: Результати розпізнавання у вигляді двійкової послідовності ($k$NN випадок)

## 5. Висновки

На основі проведених досліджень можна відзначити, що поєднання ймовірнісного та комбінаторного підходів дає можливість отримати більш коректні оцінки (що пов'язано також із їх точністю) ймовірності правильного розпізнавання за логікою їх побудови при скороченні розміру навчаючої вибірки, ніж використання лише комбінаторного підходу.

## 6. Література


[1] Капустій Б.О., Русин Б.П.,Таянов В.А. Комбінаторна Оцінка впливу зменшення інформаційного покриття класів на узагальнюючу властивість 1NN алгоритмів класифікації. Штучний інтелект.—2008.—№1.—С.49–54.

[2] Gurov, S.I. The reliability estimation of classification algorithm. Publishing department of the Computational mathematic and cybernetic faculty of Moscow State University, Moscow, 2003.

[3] Kapustii B.E., Rusyn B.P. and Tayanov V.A. Features in the design of optimal recognition systems. Automatic Control and Computer Sciences.—2008. —Vol.42.—№2.— Pp.64-70.

[4] Skurichina M., Duin R.P.W. Limited bagging, boosting and random subspace method for linear classifiers. Pattern Analysis and Applications.—2002.—no.5.—Pp.121–135.

[5] Vorontsov, K.V. Combinatorial approach to quality estimation of learning algorithms. Mathematical questions of cybernetic, 13, (2004), 5-36.