# Зважені навчаючі вибірки в розпізнаванні: формування, оптимізація, використання

*Олена В. Волченко*

Інститут інформатики і штучного інтелекту
ГВУЗ «Донецький національний технічний університет», Донецьк
lm@mail.promtele.com

**Анотація**

В роботі запропоновано загальний підхід до вирішення проблеми побудови адаптивних систем розпізнавання по зважених вибірках w-об'єктів. Досліджено особливості цих систем, що дозволило виділити ряд нових задач побудові адаптивних систем розпізнавання. Наведено та проаналізовано алгоритми формування та оптимізації зважених вибірок w-об'єктів. Для вирішення задачі класифікації запропоновано метрику для обчислення відстані між w-об'єктами та розширення відомих ефективних алгоритмів побудови вирішуючих правил класифікації на зважені вибірки. Наведено узагальнені результати тестових досліджень, що підтверджують ефективність побудови адаптивних систем розпізнавання по зважених навчаючих вибірках w-об'єктів.

## 1. Вступ

Одним з найбільш затребуваних сучасних напрямків теорії побудови систем розпізнавання є адаптивні системи, що навчаються. Інтерес до цих систем викликаний тими можливостями, які повинні забезпечувати системи цього типу, а саме [1-3]:

- адаптивність, тобто здатність змінювати свої властивості (словник ознак, навчаючу вибірку, вирішуючі правила класифікації тощо) відповідно до змін розпізнаваних об'єктів;
- робота в режимі реального часу, тобто здатність прийняти рішення про класифікацію за виділений незначний час;
- обробка класів, що значно відокремлені або суттєво перетинаються у просторі ознак;
- знаходження точних та наближених рішень, що реалізується шляхом включення до системи точних та наближених методів попередньої обробки навчаючих вибірок, побудови вирішуючих правил, прийняття рішень про класифікацію;
- наявність механізмів верифікації та валідації;
- можливість автоматичного вибору у системі мінімальної кількості параметрів для мінімізації витрат на класифікацію.

Актуальність розвитку теорії побудови адаптивних систем обумовлена, перш за все, збільшенням обсягу інформації, що обробляється системою розпізнавання на етапі проектування та функціонування. Своє застосування такі системи знаходять при побудові новинних порталів, електронних бібліотек, спам-фільтрів поштових програм, систем кредитування позичальників банків та багатьох інших [4].

Основними особливостями перелічених задач є великий обсяг початкової інформації (навчаючої вибірки), можливість її поповнення в процесі функціонування системи, необхідність виконання класифікації в режимі реального часу. Процес додавання нових об'єктів у навчаючу вибірку в більшості адаптивних систем є досить інтенсивним і відбувається на всьому протязі часу роботи системи, тому при побудові адаптивних систем розпізнавання, що навчаються виникає ряд нових невирішених задач, а саме [5]:

- скорочення розміру початкової навчаючої вибірки;
- аналіз нових об'єктів з метою визначення доцільності їх додавання до навчаючої вибірки залежно від значень їх ознак і вже наявних об'єктів в навчаючій вибірці;
- скорочення розміру навчаючої вибірки при додаванні об'єктів для зменшення часу побудови вирішуючих правил і виконання класифікації;
- коригування вирішуючих правил класифікації при додаванні нових об'єктів, що вимагає зберігання всієї навчаючої вибірки і, відповідно, може істотно збільшувати час класифікації;
- корегування словника ознак при появі нових властивостей об'єктів;
- видалення шуму та заповнення пропусків у даних.

Для побудови ефективних адаптивних систем розпізнавання у попередніх роботах [5, 7] автором запропоновано виконати перехід від традиційних до зважених навчаючих вибірок w-об'єктів, що дозволило запропонувати ряд методів вирішення зазначених задач, які у цілому стали основою побудови адаптивних систем розпізнавання, що навчаються.

У даній роботі узагальнюються та розвиваються результати вирішення задач формування, оптимізації та використання зважених навчаючих вибірок w-об'єктів у адаптивних системах розпізнавання.

## 2. Основні позначення

Зваженою навчаючою вибіркою w-об'єктів назвемо кінцеву множину $X^W = \{X_1^W, X_2^W, \ldots, X_k^W\}$. Кожен з w-

об'єктів $X_i^W$ цієї вибірки описується системою ознак $\{x_{i1}, x_{i2}, \ldots, x_{in}\} \in R^n$, тобто представляється точкою в лінійному просторі ознак, і вагою $p_i$ - цілим позитивним числом, тоді $X_i^W = \{x_{i1}, x_{i2}, \ldots, x_{in}, p_i\}$. Для кожного w-об'єкта відома його класифікація $y_i \in V$, де $V = \{V_1, \ldots, V_l\}$ - множина всіх класів системи.

У якості об'єкта, що підлягає класифікації використовуватимемо об'єкт $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sn}\}$, який описується тільки системою ознак (для однаковості присвоїмо йому одиничну вагу, тобто $p_s = 1$, тоді $X_s^W = \{x_{s1}, x_{s2}, \ldots, x_{sn}, p_s\}$).

## 3. Формування навчаючих вибірок w-об'єктів

Визначальною особливістю зважених навчаючих вибірок w-об'єктів є наявність додаткової характеристики - ваги. Аналіз широкого кола задач, що вирішуються шляхом побудови систем розпізнавання, дозволив визначити, що найчастіше у якості ваги w-об'єктів можуть використовуватися:

- значення, зумовлені особливостями досліджуваних об'єктів, або апріорно отримані від експертів;
- дані про топологічні властивості навчаючої вибірки (щільність об'єктів в деякій області простору ознак, відстань між обраними об'єктами тощо) [6, 7];
- показники впевненості класифікації w-об'єкта групою експертів [8].

Далі розглянемо ці способи визначення ваги w-об'єктів та наведемо методи обробки початкови даних.

### 3.1. Апріорне визначення ваги w-об'єктів

Вага w-об'єктів може бути задана апріорно у випадках, коли початкова вибірка включає багато об'єктів, що мають однакові значення всіх ознак, наявні дані про апріорну ймовірність появи кожного з об'єктів, експертом визначений ступень довіри до класифікації (важливість наявності у вибірці, важливість включення до вирішуючих правил тощо) навчаючих об'єктів.

### 3.2. Побудова вибірки w-об'єктів по початковій навчаючій вибірці

Попередня обробка навчаючих вибірок має на меті очистку, об'єднання та стиснення даних за умови забезпечення ефективності класифікації [1, 2]. Оскільки адаптивні системи розпізнавання характеризуються у першу чергу значним об'ємом навчаючих даних, найважливішою задачею обробки даних, на наш погляд, є їх об'єднання та стиснення.

Існуючий загальний підхід до вирішення цієї задачі, представлений, наприклад, у роботі [9], базується на відборі деякої підмножини об'єктів початкової вибірки, кожен з яких відповідає висунутим вимогам. Формування скороченої вибірки таким чином є досить ефективним, оскільки дозволяє суттєво скоротити її об'єм, однак видалення значної кількості об'єктів може призвести до зниження інформативності. Альтернативним підходом є побудова множини нових об'єктів, кожен з яких будується за інформацією про деяку підмножину об'єктів початкової вибірки та узагальнює її [2]. Основою алгоритмів даного типу є дискретизація простору ознак і аналіз отриманих частин простору незалежно один від одного. Недоліком такого підходу є відсутність аналізу кількості та розподілу об'єктів в визначених частинах, що може суттєво змінити закон розподілу значень ознак.

Пропонований підхід до побудови скороченої навчаючої вибірки [5-7] полягає у виборі множин близько розташованих об'єктів початкової вибірки, що отримали назву утворюючих множин, та заміна кожної з них одним зваженим w-об'єктом. Значення ознак w-об'єкту обчислюються за значеннями всіх об'єктів відповідної утворюючої множини. Вага є додатковою характеристикою об'єктів початкової вибірки, які були замінені одним w-об'єктом. Наведемо далі узагальнений опис алгоритмів побудови зважених навчаючих вибірок w-об'єктів.

#### 3.2.1. *Послідовний алгоритм* baseWTS

Формування w-об'єкта у базовому послідовному алгоритмі побудови зважених навчаючих вибірок baseWTS складається з трьох етапів [6].

На першому етапі виконується формування утворюючої множини, що містить деяку кількість об'єктів початкової вибірки. Побудова утворюючої множини полягає в знаходженні початкової точки (об'єкта, найбільш віддаленого від усіх об'єктів інших класів), визначенні конкуруючої точки (найближчого до початкової точки об'єкта іншого класу) і відборі до утворюючої множини таких об'єктів початкової вибірки, відстань до кожного з яких від початкової точки менше за відстань від них до конкуруючої точки.

На другому етапі виконується формування вектору ознак w-об'єкту і розрахунок його ваги. Значення ознак w-об'єкту розраховуються як координати центру мас об'єктів утворюючої множини, вага – за їх кількістю.

На третьому етапі виконується корегування початкової навчаючої вибірки - видалення об'єктів, включених до утворюючої множини.

Процес формування вибірки w-об'єктів продовжується доти, доки у початковій вибірці є об'єкти.

Зазначимо, що за аналогією з одним з найбільш відомих методів класифікації - методом k - найближчих сусідів, побудова утворюючої множини може виконуватися з однією чи k конкуруючими точками, вибір кількості яких обумовлюється розподілом об'єктів у просторі ознак.

Алгоритм baseWTS збігається, його часова складність дорівнює $O(n^2)$

#### 3.2.2. *Сітковий метод* wGridDC

Ідеєю методу wGridDC [7] є накладення сітки на простір ознак для формування множини n-мірних клітин, визначення об'єктів початкової вибірки, що належать кожній з $k$ клітин та їх заміна на w-об'єкти. Формування

об'єктів нової вибірки виконується тільки у випадку приналежності всіх об'єктів клітини до одного класу. Далі наведемо покроковий опис методу.

**Крок** 1. Формування сітки. Розраховується шаг клітини сітки $s$, який може бути однаковим для всіх ознак об'єктів чи залежати від діапазону значень конкретної ознаки. Виконується розбиття простору ознак по кожній з ознак на інтервали довжиною $s$, результатом якого є множина клітин $G$. Далі для кожного об'єкта вибірки визначається клітина, до якої належить цей об'єкт.

У результаті формування сітки та обробки об'єктів початкової навчаючої вибірки будуть сформовані непересічні підмножини об'єктів, що належать відповідним кліткам.

**Крок** 2. Формування значень ознак w-об'єктів. В залежності від особливостей розташування об'єктів у просторі ознак, можливі такі варіанти обробки вмісту клітин:

- якщо всі об'єкти клітини належать до одного класу, то значення ознак w-об'єкта розраховуються як координати центру мас об'єктів цієї клітини;
- якщо клітина не містить жодного об'єкта, то w-об'єкт нової вибірки не формується;
- якщо клітина містить об'єкти декількох класів, то вона ділиться на дві рівні за розміром клітини (по черзі вертикально або горизонтально) до тих пір, поки будь-яка з клітин усередині початкової не буде містити об'єкти тільки одного класу. Далі по кожній з цих клітин формуються об'єкти нової вибірки.

Значення ознак w-об'єктів може обчислюватись як координати центру поточної клітини, центру мас об'єктів поточної клітини або координати центру прямокутника, описаного навколо об'єктів поточної клітини.

Класифікація w-об'єкта визначається за класифікацією об'єктів, по яких він сформований.

**Крок** 3. Визначення ваги w-об'єктів. Вага w-об'єкта дорівнює кількості об'єктів початкової вибірки, що належать клітині.

В результаті виконання алгоритму буде отримана зважена навчаюча вибірка w-об'єктів.

Зазначимо, що алгоритм збігається та його часова складність дорівнює $O(k \log k)$.

### 3.3. Побудова вибірки w-об'єктів за колективною експертною класифікацією

При побудові систем розпізнавання, що навчаються, у більшості випадків класифікація об'єктів навчаючої вибірки здійснюється одним експертом і вважається вірною оскільки перевірити її правильність не представляється можливим [10]. При цьому невірна класифікація навіть незначної кількості навчаючих об'єктів може істотно змінити вирішуючи правила класифікації і привести до значного погіршення якості розпізнавання.

Вирішення цієї проблеми полягає в використанні даних про класифікацію об'єктів колективом незалежних експертів. Остаточна класифікація об'єктів визначається по результатах обробки експертних класифікацій. Якщо такі дані апріорі не можуть бути отримані, то експертами може виступати множина вирішуючих правил, побудованих по початковій вибірці.

Визначення класифікації об'єктів навчаючої вибірки за умови наявності множини експертних оцінок для них виконується шляхом розрахунку показника упевненості класифікації [8]. Його основу складає рейтинг експертів, що оцінює міру довіри класифікації об'єктів, виконаної цим експертом. Показник упевненості класифікації обчислюється як відношення сумарного рейтингу експертів, які відносять об'єкт до визначеного класу до загального рейтингу всіх експертів. Визначення класифікації кожного з об'єктів здійснюється шляхом вибору номера класу, що відповідає максимальному показнику упевненості. Вага w-об'єкту приймається рівною значенню показника упевненості класифікації класу, до якого віднесений даний об'єкт.

## 4. Оптимізація вибірок w-об'єктів

Оптимізація зважених вибірок представляє собою їх скорочення для зменшення часу класифікації та може виконуватися під час попередньої обробки вибірок для видалення шуму та під час роботи системи для оцінки необхідності додавання нових об'єктів до навчаючої вибірки.

Видалення шуму зі зваженої навчаючої вибірки виконується шляхом аналізу ваги w-об'єктів. Зважені об'єкти малої ваги при їх формуванні по початковій вибірці утворюються в області простору ознак, що вміщує малу кількість об'єктів. Тому природно припустити, що такі об'єкти утворились через помилки вимірювання їх ознак чи класифікації, тобто вони можуть бути шумом і їх видалення не погіршить ефективність класифікації. Мала вага w-об'єктів, що утворені по груповій експертній класифікації, свідчить про розбіжності в оцінці експертів і може бути хибною. При видаленні шуму з вибірки w-об'єктів встановлюється граничне значення і всі w-об'єкти, вага яких менша за це значення, не приймають участь у класифікації [11].

Оптимізація навчаючої вибірки під час додавання до неї нових навчаючих об'єктів протягом роботи системи полягає в аналізі необхідності побудови нових w-об'єктів, оскільки тривале використання системи і інтенсивне поповнення вибірки може призвести до її необмеженого зростання [6, 7].

При додаванні нових об'єктів виконується перевірка на близькість до вже наявних у вибірці w-об'єктів "свого" і "чужого" (всіх інших) класів. При обчисленні відстані враховується вага w-об'єктів. Якщо об'єкт, що додається, знаходиться ближче до w-об'єкту "свого" класу, то його додавання до вибірки в загальному випадку не внесе істотних змін у вирішуюче правило, тому створення нового w-об'єкту не виконується. Для збереження інформації про цей об'єкт виконується коригування значень ознак найближчого w-об'єкту і його вага збільшується. Якщо об'єкт, що додається, знаходиться ближче до w-об'єктів "чужого" класу, то його додавання в навчаючу вибірку дозволить скорегувати вирішуюче правило класифікації. Він оголошується новим w-об'єктом з одиничною вагою.

## 5. Класифікація об'єктів по зважених вибірках

Прийняття рішень про класифікацію у системах розпізнавання, що навчаються зазвичай виконується шляхом оцінки близькості розпізнаваного об'єкту до об'єктів навчаючої вибірки за обраною метрикою. Введення ваги для опису w-об'єктів робить неможливим використання класичних методів визначення близькості об'єктів в просторі ознак через відсутність метрики, що розраховує відстань між об'єктами, що мають неодиничну вагу. Принцип побудови w-об'єктів, згідно з яким вага w-об'єктів характеризує розташування об'єктів в просторі ознак, дозволяє запропонувати наступну метрику. Нехай кожен w-об'єкт представляється матеріальною точкою в просторі ознак і має масу, рівну вазі w-об'єкта. Тоді "близькість" двох матеріальних точок (w-об'єктів) визначається по максимальній силі тяжіння між ними. Оскільки два об'єкти прийнято вважати найближчими один до одного, якщо відстань між ними мінімальна, слід використовувати зворотну величину

$$d_W(X_i^W, X_j^W) = \frac{\sqrt{\sum_{o=1}^{n}(x_{io} - x_{jo})^2}}{p_i \cdot p_j}.$$

Класифікація за цією метрикою може здійснюватися за модифікованими методами k-найближчих сусідів [6, 8], групового урахування аргументів [12], потенційних функцій [5].

## 6. Результати експериментальних досліджень

Експериментальні дослідження виконувались на тестових та реальних даних та охоплювали всі запропоновані алгоритми побудови адаптивних систем розпізнавання по зважених вибірках. Основними з них є:

- частота невірної класифікації об'єктів тестових вибірок зменшилася в середньому на 3,6%;
- час виконання класифікації зменшився в середньому на 23%;
- перехід до зважених вибірок дозволив скоротити розмір початкових вибірок у 35-55 разів;
- додавання нових об'єктів до вибірок склало 20%, а видалення шуму зменшило частоту невірної класифікації на 0,35%.

Докладні результати досліджень наведено у роботах [5-8, 11, 12].

## 7. Висновки

У роботі запропоновано новий підхід до побудови адаптивних систем розпізнавання, що навчаються, основу якого склало використання зважених вибірок w-об'єктів. Розроблено, теоретично обґрунтовано та експериментально досліджено методи формування, оптимізації та використання вибірок w-об'єктів при вирішенні основних задач побудови систем розпізнавання, що навчаються.

## 8. Посилання


[1] Larose D.T. Discovering knowledge in Data: An Introduction to Data Mining – New Jersey, Wiley & Sons, 2005. – 224 p.

[2] Pal S.K., Mitra P. Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing – Chapman and Hall, 2004. – 280 p.

[3] Olson D.L. Advanced Data Mining Techniques / D.L. Olson, D. Delen – Springer-Verlag Berlin, 2008. – 180 p.

[4] Cherkassky V., Mulier F. Learning from data. Concept, theory and methods, 2nd ed. – New Jersey: John Wiley & Sons, 2007. – 540 p.

[5] Розробка теоретичних засад і методів реалізації відкритих систем автоматичного розпізнавання, що навчаються: способи оптимізації навчаючих вибірок і методи побудови зважених вирішуючих правил класифікації: звіт з НДР: Тема GP/F32/130, Грант Президента України для підтримки наукових досліджень молодих учених на 2011 р. / керівник О.В. Волченко. – Донецьк, ДВНЗ «ДонНТУ», 2011. – 67 с.

[6] Волченко Е.В. Метод построения взвешенных обучающих выборок в открытых системах распознавания // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов (ММРО-14)», Суздаль, 2009. – М.: Макс-Пресс, 2009. – С. 100 – 104.

[7] Волченко Е.В. Сеточный подход к построению взвешенных обучающих выборок w-объектов в адаптивных системах распознавания // Вісник Національного технічного університету "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ", 2011. – № 36. – С. 12 – 22.

[8] Волченко Е.В. Построение обучающей выборки w-объектов на основе коллективного решения группы экспертов // Штучний інтелект. – 2011. – №1. – С. 147 – 153.

[9] Загоруйко Н.Г. Прикладные методы анализа знаний и данных – Новосибирск: Издательство института математики, 1999. – 270 с.

[10] Лапко А.В., Лапко В.А., Ченцов С.В. Непараметрические модели распознавания образов в условиях малых выборок // Автометрия. – 1999. – № 6. – С. 105 – 113.

[11] Volchenko E.V. Research of features in association of training sample objects to meta-objects // 9th International Conference on "Pattern recognition and image analysis: new information technologies": Conference Proceeding, 2008. – Vol. 2. – P. 291-294.

[12] Волченко Е.В. Расширение метода группового учета аргументов на взвешенные обучающие выборки w-объектов // Вісник Національного технічного університету "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ", 2010. – № 31. – С. 49 – 57.