УДК 004.93

ОЦІНКА ЕФЕКТИВНОСТІ МОДЕЛІ НАВЧАННЯ ТА ЯКОСТІ РОБОТИ МЕТРИЧНИХ КЛАСИФІКАТОРІВ

© Капустій Б.О., Русин Б.П., Таянов В.А.

Національний університет «Львівська політехніка»
кафедра теоретичної радіотехніки та радіовимірювань ІТРЕ
вул. С. Бандери, 12, м. Львів, 79013, Україна

Фізико-механічний інститут ім. Г.В. Карпенка НАН України
відділ методів та систем обробки, аналізу та ідентифікації зображень
вул. Наукова, 5, м. Львів, 79601, Україна
e-mail: *vtayanov@ipm.lviv.ua*

Abstract. In this paper the full conception of the probabilistically combinatorial approach has been presented. This conception is the result of previous long preliminary works. The approach gives the possibility to establish the reasons of recognition algorithms overtraining, to define the possible ways of it reduction and to build the most precise estimates of the recognition probability. The combinatorial approach works with determined data of the recognition process and the probabilistic one determines the probability of these results existence. The most usefulness of the combinatorial approach consists in the possibility to determine the training set variation influence on the different algorithms and select the most appropriate one from these algorithms or algorithm composition. The probabilistic part of this approach determines the probability of results obtained on the basis of combinatorial approach.

Вступ

Усі класифікуючі алгоритми можуть бути поділені на три групи: алгоритми з навчанням, із самонавчанням та алгоритми, що не використовують навчання як такого. Найбільш важливими і цікавими є алгоритми, що використовують навчання. Ці алгоритми є об'єктом дослідження в рамках теорії машинного навчання (Theory of Machine Learning), яка доволі швидко й успішно розвивається на протязі останніх десяти років [7]. У рамках цієї теорії розглядаються такі важливі питання, як визначення оптимального складу навчаючої вибірки, навчання класифікаторів та побудова оптимальної композиції класифікаторів, що задовольняє певним умовам, а також генерація та селекція 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 оптимального складу навчаючої вибірки та набору найбільш інформативних ознак), так і надлишковості (складності) самого апарату класифікації, а, власне, 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 алгоритмів і досягти найбільшого значення ймовірності правильного розпізнавання.

У даній роботі розглядаються метричні класифікатори. Серед усіх метричних класифікаторів найбільш часто для побудови практичних цільових систем, що застосовуються в різних галузях діяльності людини, застосовуються класифікатори

типу kNN, які використовують ідею класифікації на основі найближчого сусідства. Переваги простих метричних алгоритмів типу kNN є такими:

- *Простота реалізації та можливість введення різноманітних модифікацій;*
- *Можливість інтерпретації розпізнавання шляхом пред'явлення користувачу найближчого об'єкта або декількох. «Прецедентна» логіка роботи алгоритму є добре зрозумілою експертам з таких предметних областей, як медицина, біометрія, юриспруденція, металофізика, робототехніка та ін.*

1. Проблеми теорії машинного навчання класифікуючих алгоритмів

В сучасній теорії машинного навчання існують дві серйозні проблеми: отримання точних верхніх оцінок ймовірності такого негативного явища, як перенавчання, та способів боротьби з ним. На даний момент найбільш точні з відомих оцінок значно завищені. Експериментально вдалося встановити основні причини завищення оцінок. У порядку зменшення впливу, вони є наступними:

1. *Нехтування ефектом розшарування або локалізації сімейства алгоритмів. Дана проблема обумовлюється тим, що залежно від виду задачі використовується не все сімейство алгоритмів, а лише певна його частина. Коефіцієнт завищеності знаходиться в межах від декількох десятків до сотень тисяч.*
2. *Нехтування схожістю алгоритмів. Коефіцієнт завищеності становить для цього фактора від декількох сотень до десятків тисяч. Цей фактор завжди присутній і менш залежний від виду задачі, ніж перший.*
3. *Експоненційна апроксимація «хвоста» гіпергеометричного розподілу. В цьому випадку коефіцієнт завищеності може складати декілька десятків.*
4. *Верхня оцінка профіля різноманітності представляється одним скалярним коефіцієнтом різноманітності. Коефіцієнт завищеності часто порядку одиниці, однак у деяких випадках може досягати декількох десятків.*

Причина ефекту перенавчання обумовлюється тим, що використовуються алгоритми з мінімальним числом помилок на навчаючій вибірці, тобто відбувається однобічне налаштування цих алгоритмів. Перенавчання буде тим більшим, чим більшу композицію з алгоритмів ми використовуємо для класифікації, якщо алгоритми беруться з розподілу випадково і незалежно. У випадку залежності алгоритмів (в реальній ситуації вони, як правило, такими і є) перенавчання зменшиться. Отже, при виборі навіть одного з двох алгоритмів може виникнути перенавчання. Розшарування алгоритмів за числом помилок та збільшення їхньої подібності зменшують ймовірність перенавчання. Розглянемо для прикладу дуплет «вибірка-алгоритм». Кожний алгоритм покриває певну частину об'єктів навчаючої вибірки. Якщо використовувати внутрішні критерії [6] (наприклад, у випадку метричних класифікаторів), то можна оцінити стійкість цього покриття і звузити число покритих об'єктів згідно із заданим рівнем стійкості. Таким чином, для того щоб покрити більшу кількість, потрібно застосувати більшу кількість алгоритмів. Ці алгоритми мають бути схожими і мати різний рівень помилок. Однак, якщо використовуються тестові дані,

до яких композиція алгоритмів неадаптована, то помилка класифікації може досить сильно відрізнятися від мінімальної, отриманої на навчаючих даних. З іншого боку, цікавою представляється задача по визначенню кількості надлишкової інформації у навчаючих даних. Доцільність у зменшенні навчаючих даних полягає в тому, що для кожного конкретного випадку зменшується також і кількість об'єктів інших класів, що заважають класифікації. При цьому потрібно оцінити середнє значення розміру класу, що забезпечує потрібний рівень частоти помилок. Зменшення кількості навчаючих даних також означає зменшення розміру класів на етапі тестування. Оцінка ефекту від пониження розміру навчаючих даних дає можливість визначити структуру цих даних, тобто співвідношення між еталонними об'єктами та об'єктами-викидами, пороговими або неінформативними. Крім того, чим менший розмір класу, тим менший час, потрібний для прийняття рішення. Однак найбільшою цінністю даного підходу є те, що він дозволяє детальніше вивчити і глибше зрозуміти явище перенавчання алгоритмів.

2. Підходи до оцінки якості роботи класифікаторів

Якість роботи класифікаторів прийнято характеризувати через поняття відступу (margin), що представляє відстань об'єкта від розділювальної гіперплощини. Чим більший відступ, тим кращим вважається класифікатор. Однак якщо всі об'єкти або переважна їх більшість мають приблизно однаковий 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. Такий підхід породжує одну з основних причин, що обумовлюють ефект перенавчання. Однобічне налаштування алгоритма розпізнавання на основі близької за суттю навчаючої інформації приводить до того, що на контрольній вибірці він може часто помилятись, навіть якщо він не помилявся на навчаючій вибірці. Дійсно, ймовірність того, що в умовах навчаючої вибірки зустрінеться така ж ситуація, є близькою до нуля.

Тому для навчання прийнято використовувати несхожі і «важкі» для алгоритма об'єкти з малими значеннями відступу. Ця ідея використана, зокрема, у методі опорних векторів (Support Vector Machine) або методі зваженого голосування. Застосуємо узагальнений підхід для характеристики класифікаторів на основі поняття відступу. Результатом роботи метричних класифікаторів є ранжовані дані (посортовані за ступенем подібності до тестового об'єкти бази даних). Для таких класифікаторів поняття відступу представляється наступним чином. Вводиться еквівалентна до класичного відступу характеристика, яка для даного об'єкта може бути представлена як відносна відстань між його відстанями від тестового об'єкта та від усередненого об'єкта бази даних або останнього об'єкта з однорідної (стратегічної) [3] 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. **Характеристика метричних класифікаторів.** Для більш строгого означення даної характеристики потрібно ввести поняття розподілу відстаней між об'єктами. Оскільки значення відстаней може бути довільним, то процедура непараметричного оцінювання розподілу неусіченими ядерними функціями буде коректною. Якщо оцінене математичне сподівання нормального закону розподілу рівне $\widehat{\mu}$, а дисперсія – $\widehat{\sigma^2}$, то відносна відстань може бути оцінена через параметр z у вигляді $\widehat{z} = \frac{x-\widehat{\mu}}{\widehat{\sigma}}$. Тоді нормальний закон розподілу відстаней представляється як $p(\widehat{z}) = \frac{1}{2\pi} e^{-\frac{\widehat{z}^2}{2}}$. На практиці більшу користь має не саме значення параметру $\widehat{z}$, а функція розподілу $P(\widehat{z}) = \int_{-\infty}^{\widehat{z}} p(\widehat{z}) d\widehat{z}$. Функція розподілу в даному випадку є однозначною характеристикою відокремлення «свого» об'єкта від сукупності «чужих» об'єктів. Оскільки гіперплощина у випадку порогових класифікаторів виконує роль границі між класами, то еквівалентна їй характеристика для метричних класифікаторів визначає, наскільки добре об'єкти «свого» класу відокремлюються від сукупності «чужих» об'єктів. Ця характеристика має строге математичне обґрунтування і є функцією розподілу ймовірностей [6].

Задача збільшення відступу у випадку метричних класифікаторів на основі навчаючої вибірки вирішується наступним чином. Першим стратегічним напрямком є зменшення дисперсії густини розподілу відстаней між об'єктами, а також збільшення середнього значення цього розподілу. В рамках теорії послідовного аналізу це означає, що може бути збільшена база даних, а ймовірність правильного розпізнавання залишатиметься на тому самому рівні. Другий стратегічний напрямок полягає в тому, що потрібно прагнути, щоб розподіл відстаней був якомога ближчим до нормального. Ця ідея обґрунтовується наступним чином.

Розглянемо розподіл ознак у лінійному багатовимірному або відстаней між об'єктами в одновимірному просторі та проведемо його аналіз. Ймовірність помилки розпізнавання для $\mu = 0$ може бути представлена як $\int\limits_{|x|\geq\theta} p(x)\,dx$, де θ – поріг. Згідно з нерівністю Чебишева [5] отримаємо $\int\limits_{|x|\geq\theta} p(x)\,dx \leq \frac{\sigma^2}{\theta^2}$. Розглянемо випадок рівності середніх значень та дисперсій розподілу $p(x)$. Верхня межа для одномодального розподілу з модою μ_0 обчислюється за допомогою нерівності Гауса наступним чином [4]:

$$P\Big(|x-\mu| \geq \lambda\tau\Big) \leq \frac{4}{9\lambda^2}, \tag{1}$$

де $\tau^2 \equiv \sigma^2 + (\mu - \mu_0)^2$.

Нехай $\mu = \mu_0 = 0$ та $\tau \equiv \sigma$, тоді поріг $\theta = \lambda\sigma$, а $\lambda = \frac{\theta}{\sigma}$. Таким чином, нерівність Гауса для порогу θ може бути представлена у вигляді

$$\int\limits_{|x|\geq\theta} p(x)\,dx \leq \frac{4\sigma^2}{9\theta^2}. \tag{2}$$

Як видно з (2), оцінка Гауса зверху для одномодального закону розподілу є в 2.25 разів кращою, ніж оцінка Чебишева для довільних розподілів, що підтверджує

суттєвий вплив виду розподілу ознак на ймовірність правильної класифікації. Нормальний закон розподілу ймовірностей має однакові моду, медіану та математичне сподівання. Крім цього, на практиці цей закон є одним із найпоширеніших. З іншого боку, нормальний закон розподілу характеризується максимальним значенням ентропії при однакових значеннях перших моментів. А це означає, що отримується мінімальна помилка класифікації для нормально розподілених класів.

Розглянемо способи обчислення відстаней. Один із способів полягає в застосуванні різних метрик, серед яких у першу чергу можна відзначити узагальнену метрику Мінковського та косинусну метрику. Інший спосіб обчислення відстаней передбачає використання ядерних функцій. Найчастіше вживаними ядерними функціями є три. Це – радіальна базисна функція, сигмоїдальна та поліноміальна функції. Найбільш поширеною і вживаною серед них є радіальна базисна функція. Спільною рисою обох способів обчислення відстаней є використання зважених ознак, що є головною задачею, яку вирішує той чи інший метод обчислення відстаней. Зважування ознак дозволяє коректувати напрям гіперплощини в гіперпросторі таким чином, щоб найбільш оптимально розділяти класи. Для певного набору ознак вибираються такі ваги, які для переважної більшості об'єктів є оптимальними.

Розглянемо, як приклад, представлення мір відстаней між векторами ознак $\mathbf{x}$ та $\mathbf{y}$ через міру Манхетена – просту лінійну міру із зваженими коефіцієнтами a_i :

$$d(x,y) = \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{3}$$

де $d(x,y)$ — довільна міра відстаней між векторами $\mathbf{x}$ та $\mathbf{y}$

Міру відстаней Мінковського, як найбільш узагальнену міру, що використовується в теорії розпізнавання образів, можна представити у вигляді

$$d(x,y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^p \Big)^{\frac{1}{p}} = C(p) \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{4}$$

де $C(p) = \Big(\sum_{i=1}^{n} a_i |x_i - y_i| \Big)^{\frac{1-p}{p}}$; $a_i = (|x_i - y_i|)^{p-1}$; $p > 0$.

Таким же чином визначаються коефіцієнти в косинусній метриці, метриці Канбера та інших метриках для класифікаторів типу kNN, а також параметри трьох згаданих ядерних функцій для інших типів класифікаторів. Ці задачі вирішуються на основі конкретної навчаючої вибірки. Таким чином, проблема полягає не у виборі найбільш оптимальної метрики, а у визначенні ваг ознак для того чи іншого конкретного випадку. Один із способів обчислення ваг є використання функцій відстані у вигляді метрик або ядерних функцій. Цей спосіб є найбільш простим, математично добре обґрунтованим та зрозумілим. Вагові коефіцієнти ознак дискретно згортаються з певним видом функцій (наприклад, ядерними функціями), що в результаті дає відстань. Якщо використовуються ядерні функції, то метрика буде результатом непараметричного оцінювання відстані між двома векторами ознак. Точність оцінювання оптимальної відстані буде визначатися кількістю та набором ознак. Це підтверджує,

наприклад, порівняння результатів непараметричного оцінювання густини розподілу на основі класичного методу вікна Парзена та методу опорних векторів В. Вапніка [8, 9], в якому використовується процедура оптимізації шляхом розв'язку задачі квадратичного програмування. Хоч непараметричне оцінювання відбувається за допомогою лише невеликої частини опорних векторів, однак результати оцінювання є більш точними від методу Парзена. Звідси випливає, що використання порогових і метричних класифікаторів є абсолютно еквівалентним, а задача полягає лише у знаходженні відповідних параметрів, що максимізують (мінімізують) той чи інший функціонал штрафу за помилку класифікації.

2.2. **Аналіз процесу класифікації при використанні метричних класифікаторів.** Під метричним класифікатором розуміють відображення виду

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

Дія такого класифікатора проявляється в тому, 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, є те, що рішення, прийняте цими класифікаторами, не залежить від порогу. Разом з тим метричні класифікатори мають достатню кількість ступенів свободи для їх налаштування і є, як правило, більш стійкими до впливу зовнішніх факторів, ніж порогові класифікатори, за рахунок їх інтегрального характеру. Серед метричних класифікаторів за ступенем збільшення складності можна виділити 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 $(\frac{w_i}{w_{i+1}})$ у 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 kNN алгоритм, при $1 < \frac{w_i}{w_{i+1}} < 2$ – зважений kNN алгоритм, 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 результаті представляється 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 і обчислюється за допомогою ядерних функцій із постійним або змінним вікном h_i, а центр ядра знаходиться в самому класифікованому об'єкті. Оскільки у методі Парзена ваги об'єктів 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» у 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 в силу оптимальності повинен бути нормальним [6]), то обидва підходи (класичний та ймовірнісно-комбінаторний) є абсолютно еквівалентні щодо задачі класифікації. Принципова відмінність полягає в тому, що параметри ядер у ймовірнісно-комбінаторному підході визначаються на основі навчаючої вибірки, а також є функціями процесу відбору ознак, способу обчислення відстаней тощо.

3. Суть імовірнісно-комбінаторного підходу

Основна мета поєднання двох підходів полягає в тому, щоб досягнути більшої точності та 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 «своїм» образам, а 0 – «чужим». Приклад такої послідовності показаний на рис. 1.

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

Рис. 1. Результати розпізнавання у вигляді двійкової послідовності (kNN алгоритм)

Розглянемо випадок $ENT(\frac{k}{2})+1 \le s^*$. Визначимо ймовірності того, що серед послідовності образів «свого» класу заданої довжини s 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 kNN класифікатора є добутком цих імовірностей. Визначимо ймовірність помилкової класифікації, обумовленої образами з «чужих» класів:

$$q_j = \frac{1}{C_s^{s^*}} \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_{i,j} m_{i+j-1}|} C^{s^*-j}_{s-|\bigcup_{i,j} m_{i+j-1}|}; |\bigcup_{i,j} m_{i+j-1}| > ENT(\frac{k}{2})+1. \quad (5)$$

Обчислимо довірчу ймовірність для довільної послідовності з образів «свого» класу:

$$P_(q_j) = \frac{1}{C_s^{s^*}} \sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i \ell_i|} C^{s^*-j}_{s-|\bigcup_i \ell_i|}. \quad (6)$$

Ймовірність правильного розпізнавання при застосуванні kNN класифікатора визначається добутком імовірності (6) та доповнення до ймовірності (5):

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

$$-\frac{1}{(C_s^{s^*})^2}\Big(\sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_i \ell_i|} C^{s^*-j}_{s-|\bigcup_i \ell_i|}\Big)\Big(\sum_{j=ENT(\frac{k}{2})+1}^{s^*} C^j_{|\bigcup_{i,j} m_{i+j-1}|} C^{s^*-j}_{s-|\bigcup_{i,j} m_{i+j-1}|}\Big). \quad (7)$$

Роль імовірнісної частини в імовірнісно-комбінаторному 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. Це означає, що 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 від ваги наступного старшого розряду, тобто заміщення довільного образа «свого» класу у списку еквівалентне почерговому заміщенню всіх 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 в послідовності образів зі «свого» класу образами з «чужих» класів. З іншого боку, довільні ваги можна виразити через показник ступеня 2, що також спрощує представлення та обчислення цих імовірностей. Таким чином, ймовірність існування однорідної послідовності з образів «свого» класу обчислюється на основі розподілу відстаней і є функцією від параметрів алгоритму розпізнавання. Приймається така послідовність, для якої ймовірність існування є достатньою.

Відтак застосовується комбінаторна частина підходу, яка дозволяє обчислити ступінь впливу пониження розміру класів на ймовірність правильного розпізнавання. Оскільки ймовірнісна частина підходу визначається параметрами алгоритму розпізнавання, то поєднання ймовірнісної та комбінаторної частин дозволяє більш точно описати ефект від зменшення кількості навчаючих даних.

Наприкінці розглянемо покроково приклад швидкого обчислення ймовірності заміщення «свого» образу з послідовності, де співвідношення між вагами об'єктів є цілий ступінь числа 2. Нехай, 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 ваги об'єкта «свого» класу. Знайдемо ймовірність заміщення об'єкта з вагою 2^9 порівняно із об'єктом 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 λ 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.

Оскільки такий параметр, як кількість найближчих сусідів, визначає надійність роботи kNN метричних класифікаторів як в імовірнісній частині запропонованого підходу, так і в комбінаторній, то він дає можливість визначити різницю між різними алгоритмами розпізнавання на основі методу kNN. Ця різниця має ймовірнісний характер і, як очікується, несе більше інформації для прогнозування ефекту перенавчання, ніж відомі підходи [1].

Висновки

На основі проведених досліджень встановлено, що поєднання ймовірнісного та комбінаторного підходів дає можливість отримати більш коректні оцінки ймовірності правильного розпізнавання за логікою їх побудови при скороченні розміру навчаючої вибірки, ніж використання лише комбінаторного підходу.

Список литературы

1. *Воронцов К.В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов.—М.: Физматлит, 2004. — Т. 13. — С. 5–36.
2. *Гуров, С.И.* Оценка надежности классифицирующих алгоритмов.— М.: Издательский отдел ф-та ВМиК МГУ, 2002. — 45с.
3. *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.
4. Математическая Энциклопелия: Гл. ред. И.М. Виноградов, т. 5. Слу—Я—М., «Советская Энциклопедия», 1984.—1248 стб., ил.
5. *Шлезингер М., Главач В.* Десять лекций по статистическому и структурному распознаванию—Киев: Наукова думка, 2004.—545 с.
6. *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.
7. *Skurichina M., Duin R.P.W.* Limited bagging, boosting and random subspace method for linear classifiers. Pattern Analysis and Applications.—2002.—№5.—Pp.121-135.
8. *Vapnik V.* The nature of statistical learning theory.—2 edition.—Springer-Verlag, New York, 2000.
9. *Webb A.* Statistical Pattern Recognition, John Wiley and Sons Inc, 2nd ed., New York, 2002.

Статья поступила в редакцию 14.01.2009