Нерівності для моделей узагальнених метричних класифікаторів

Віталій Таянов

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

Abstract

В даній роботі розглядаються оцінки ймовірності коректної роботи статистичних алгоритмів розпізнавання типу kNN. Визначено оцінки верхньої та нижньої межі ймовірності розпізнавання, використовуючи нерівність Гауса і властивості нормального закону розподілу. Можливість застосування нерівності Гауса обумовлена специфікою розподілу відстаней між об'єктами при використанні різного роду метрик. На основі непараметрично оцінених розподілів відстаней між образами в межах кожного з класів запропоновано модель класифікатора, що може працювати навіть в умовах нульових відстаней між об'єктами.

1. Вступ

На сьогоднішній день методи оцінки ймовірності правильного розпізнавання базуються на алгоритмах ковзаючого контролю (cross validation) [1]. Однак такі алгоритми ковзаючого контролю, як виключення по одному та ряд інших, вимагають великої кількості обчислень і комбінаторних перегрупувань вибірки. Тому потрібно розробити підходи щодо побудови верхніх оцінок для значно меншої кількості комбінаторних перегрупувань. Це можливо зробити тому, що навчаючі дані практично завжди містять надлишок інформації, який проявляється в її частковому дублюванні. З точки зору перенавчання побудова верхніх оцінок означає, що перед алгоритмом класифікації поставлено найбільш складну задачу (подано найбільш складну з точки зору класифікації вибірку), яка включає в себе довільні більш прості. Тобто, верхні оцінки ймовірності правильного розпізнавання моделюють класифікацію найбільш незручних об'єктів навчаючої вибірки. Якщо ймовірність того, що знайдуться більш несприятливі об'єкти на контрольній вибірці мала, то доцільно говорити про те, що доповнення до цієї ймовірності визначатиме надійність оцінок. Поряд з тим, що будуються верхні оцінки для всієї вибірки або для сукупності підвибірок, також оцінюється зверху ймовірнісна стійкість покриття кожного об'єкта зокрема. Таким чином отримується точніша повна оцінка зверху для ймовірності правильного розпізнавання.

З іншого боку, якщо проводити непараметричне оцінювання розподілу вектору відстаней від тестового об'єкта до об'єктів навчаючої вибірки, то звичайний kNN алгоритм набуває властивостей зваженого kNN алгоритму та оптимальних властивостей баєсівського алгоритму. Вага впливу кожного об'єкта на результати класифікації визначається через ядерні функції, за допомогою яких робиться непараметричне оцінювання. Неперервні розподіли, отримані непараметричним оцінюванням, повністю відповідають баєсівській моделі класифікатора. Якщо в баєсівських моделях використовувати однопороговий критерій Неймана-Пірсона чи Зігерта [2, 3], то це відповідає моделі класифікатора з використанням лінійних гіперплощин. Якщо ж використати два пороги, як у статистичній моделі Вальда [4], то це відповідає лінійному алгоритму опорних векторів, відомому у зарубіжній літературі як SVM (Support Vector Machine)[5]. Таким чином, як бачимо, всі основні моделі класифікації є частковими випадками узагальненої моделі метричного класифікатора.

Якщо розглядати критерії оптимізації оцінок якості і достовірності роботи класифікаторів, то тут можливі два підходи. Один підхід передбачає максимізацію математичного сподівання оцінки, а другий–мінімізацію її дисперсії. Пропонується в якості критерію використати відношення математичного сподівання оцінки до її дисперсії. В [6] такі відношення названі z–параметрами.

2. Побудова оцінок ймовірності коректної роботи алгоритмів типу kNN

Якість роботи класифікаторів, що будуються на основі рангового голосування та з використанням розділювальних гіперплощин (R моделей [7]) 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 головних причин, що обумовлюють ефект перенавчання. Однобічне налаштування алгоритма на основі близької за суттю навчаючої інформації призводить до того, що на контрольній вибірці він може часто помилятись, навіть якщо 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 алгоритму 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)[8] послідовності "своїх"об'єктів. Передбачається, що хоча б частина "своїх"об'єктів розташовуються на початку списку можливих претендентів. Таким чином, гарантується коректність даного означення. Для більш строгого означення даної характеристики потрібно ввести поняття розподілу відстаней між об'єктами. Оскільки значення відстаней може бути довільним за абсолютною величиною, то процедура непараметричного оцінювання розподілу неусіченими ядерними функціями буде коректною.

Нехай непараметрично оцінена густина розподілу відстаней між об'єктами, заданих векторами $\mathbf{x}$ та $\mathbf{y}$: $p(x)$, $x \to d(\mathbf{x}, \mathbf{y})$. Згідно з нерівністю Чебишева [3] ймовірність того, що знайдеться відстань, яка перевищить деяке порогове значення відстаней θ, рівна $\int\limits_{|x|\geq\theta} p(x)dx \leq \frac{\sigma^2}{\theta^2}$.

Розглянемо випадок рівності математичного сподівання та моди розподілу $p(x)$. Верхня межа одномодального розподілу з модою $\mu = 0$ за допомогою нерівності Гауса[9] представляється у вигляді:

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

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

Нехай $\mu = \mu_0 = 0$ і $\tau \equiv \sigma$. Тоді поріг $\theta = \lambda\tau = \lambda\sigma$, а $\lambda = \frac{\theta}{\sigma}$. Отже, нерівність Гауса для порогу θ може бути представлена у вигляді:

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

Таким чином, згідно з нерівністю Гауса для одномодальних розподілів з модою, рівною математичному сподіванню, оцінка є в 2.25 разів кращою за ту, яка отримується за допомогою нерівності Чебишева. Це є максимально хороша оцінка за умови, що невідомий конкретний вид розподілу, а відомі лише певні його властивості. Необхідною і достатньою умовою рівності моди математичному сподіванню є симетричність одномодального розподілу. Однак у загальному випадку реальний закон розподілу не є симетричним. При цьому можливі ліва або права асиметрії функції густини розподілу ймовірностей (ФГРЙ).

3. Нерівності для двох класів розподілів відстаней між об'єктами

Розділимо ФГРЙ на дві частини, а саме: частину, що знаходиться справа та зліва від максимума. Якщо права частина ФГРЙ більша за ліву, то вважається, що це права асиметрія, а якщо навпаки, то ліва (рис.1,2). Розглянемо оцінки за допомогою нерівності Гауса для обидвох випадків. У випадку правої асиметрії застосуємо наступний прийом. Зробимо розподіл симетричним за лівою частиною, тобто ліву частину залишаємо незмінною та відображаємо її симетрично відносно максимума замість вихідної правої частини. Такий прийом застосовано лише для оцінювання еквівалентної дисперсії і з питаннями нормування розподілу до одиничної площі він не має нічого спільного. Нехай деяка точка x_0 належить лівій частині розподілу. Тоді еквівалентна функція розподілу ймовірностей (ФРЙ) $P(X < x_0)$ для симетричного випадку (в даному випадку функція нормована до одиничної площі) буде завжди більшою від вихідного значення у кожній точці лівої частини розподілу. Відзначимо, що нас цікавлять перші об'єкти у списку можливих претендентів, що відповідають лівій частині розподілу. Тоді ФРЙ буде верхньою оцінкою для помилки розпізнавання.

Проаналізуємо отриманий результат. Фактичне зменшення площі під кривою ФГРЙ означає, що на результати розпізнавання і прийняття рішення впливають не всі об'єкти, а лише їх части-

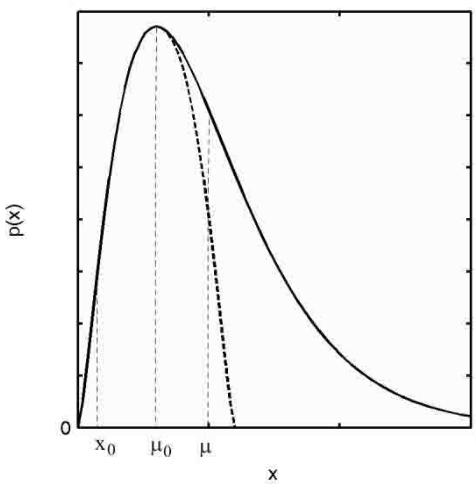

Рис. 1: Права асиметрія в ФПРВ.

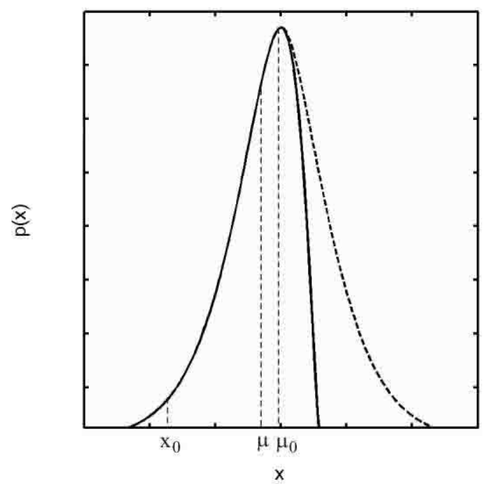

Рис. 2: Ліва асиметрія в ФПРВ.

на, що відповідає реальній ситуації. До того ж інтерес представляють відхилення відстаней вліво від математичного сподіваяння при використанні kNN класифікаторів з невеликими значеннями k. Оскільки оцінка дисперсії ФГРЙ для побудови оцінки Гауса проводилась за лівою частиною, то очевидно, що ця оцінка у випадку симетричної ФГРЙ буде меншою за вихідну, що робить її більш точною. До того ж симетрія дозволяє зробити оцінку Гауса максимально точною згідно з нерівністю (2), а все разом дозволяє суттєво покращити загальну верхню оцінку.

Розглянемо ФГРЙ у випадку лівої асиметрії. При цьому єдино можливим є лише симетричне відображення лівої частини у праву, оскільки лише тоді можлива коректна верхня оцінка. Тепер дисперсія симетричної ФГРЙ буде більшою, ніж вихідної, а єдиною перевагою такого перетворення буде симетричність щойно отриманого закону розподілу. У даному випадку також немає необхідності у нормуванні ФГРЙ при оцінюванні еквівалентної дисперсії. Збільшення площі під кривою означає, що включені додаткові об'єкти, які не приймають участі у розпізнаванні. Це погіршує оцінку Гауса, оскільки збільшується значення оціненої дисперсії. Рішення про те, яку оцінку використовувати - з перетворенням симетрії або по вихідному розподілу необхідно приймати, маючи значення математичного сподівання, моди та дисперсії обидвох розподілів. Проаналізуємо зв'язок оцінки Гауса зі значенням ФРЙ $P(X < x_0)$. Права частина ФГРЙ не представляє інтересу, тому якщо замість оцінки Гауса взяти ФРЙ, то це буде оцінкою зверху по відношенню до самої оцінки. При цьому не мають значення ані вид асиметрії, ані сама асиметрія в законі розподілу взагалі. Тобто, верхня оцінка значеннями ФРЙ по відношенню до оцінки Гауса стосується як симетричних, так і асиметричних законів розподілу. Таким чином, завищеність оцінки Гауса по відношенню до значень ФРЙ компенсується лише у випадку правої асиметрії. У випадку лівої асиметрії ступінь компенсації залежить від співвідношення між значенням дисперсії та різниці $|\mu - \mu_0|$. Якщо ФГРЙ не має чітко вираженої структури (існування максимуму, симетрія), тоді можна скористатися непараметричним оцінюванням, в результаті якого отримується неперервна ФГРЙ. Цю функцію можна інтегрувати та диференціювати. З іншого боку, при збільшенні розміру чужого класу та незмінному розмірі свого зменшується ймовірність правильного розпізнавання. Тому потрібно по можливості забезпечувати рівність розмірів класів, щоб не було перекосу в пріоритетах. Якщо ж цього не вдасться зробити, то при обчисленні слід враховувати ймовірність появи об'єктів того чи іншого класу.

Оскільки нормальна ФГРЙ характеризується мінімальною помилкою класифікації для даного порогу θ і не перевищує $\frac{4\sigma^2}{9\theta^2}$ у випадку одномодальної симетричної ФГРЙ або ФГРЙ з правою асиметрією, то двостороння нерівність для даної помилки розпізнавання ε може бути записана у вигляді:

$$0.5(1 - erf(\frac{\theta}{\sigma})) \leq \varepsilon \leq \frac{4\sigma^2}{9\theta^2}, \tag{3}$$

де $\mu = 0$.

Проаналізуємо можливу загальну форму потенційно отримуваних ФГРЙ відстаней між об'єктами. Всі розподіли матимуть максимуми, оскільки функція ФГРЙ існує на інтервалі $[0, \infty)$, а густина в околі 0 та для великих відстаней не може бути значною, тому що такі події малоймовірні. Права асиметрія є набагато більш імовірною, оскільки розподіл відстаней обмежений з лівого боку нулем, а з правого боку він не має строгих обмежень.

4. Оцінки ймовірності класифікації алгоритмами kNN в умовах двох класів

Розглянемо поширену задачу класифікації в умовах двох класів. Позначимо розміри класів як s_1 та s_2. Тоді, якщо ймовірність заміщення об'єкта з класу розміром s_1 у межах довірчого інтервалу рівна ε_1, то ймовірність незаміщення об'єктів із цього класу об'єктами з класу розміром s_2 рівна $(1-\varepsilon_1)^{s_2}$ при умові незалежності об'єктів [6]. Для іншого класу при відповідних змінах у позначеннях ця ймовірність дорівнюватиме $(1-\varepsilon_2)^{s_1}$. Якщо тепер ввести деякий віртуальний клас і вважати, що заміщення якогось об'єкта цього класу об'єктами із згаданих двох класів є вірогідною подією, то можна записати наступне рівняння:

$$\gamma((1-\varepsilon_1)^{s_2} + (1-\varepsilon_2)^{s_1}) = 1, \tag{4}$$

звідки множник пропорційності γ знаходиться тривіально. Часом зустрічаються ситуації, коли відстані між об'єктами рівні 0. Нехай густини розподілів у нульовій точці рівні $p_1(0)$ та $p_2(0)$. Оцінка співвідношення між імовірностями може бути задана у вигляді $\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$або $\ln\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$. При цьому потрібно зробити граничний перехід від ФРЙ до ФГРЙ, оскільки вони пов'язані між собою операцією диференціювання. Співвідношення $\ln\frac{p_1(0)^{s_2}}{p_2(0)^{s_1}}$ ($\ln\frac{p_2(0)^{s_1}}{p_1(0)^{s_2}}$) або в загальному $\ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}}$ ($\ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}}$) можна використати для побудови класифікатора виду

$$\begin{aligned} \ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} &> \gamma_1; \\ \ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} &< \gamma_1, \end{aligned} \tag{5}$$

або

$$\begin{aligned} \ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} &> \gamma_2; \\ \ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} &< \gamma_2, \end{aligned} \tag{6}$$

де значення $\ln\frac{p_1(\theta)^{s_2}}{p_2(\theta)^{s_1}} = 0$ або $\ln\frac{p_2(\theta)^{s_1}}{p_1(\theta)^{s_2}} = 0$ не впливають на результати класифікації і рішення може бути прийняте на користь довільного класу. У випадку непараметричного оцінювання ймовірність такого значення практично рівна 0.

5. Acknowledgements

В роботі побудовано та досліджено оцінки ймовірності правильної класифікації для класифікаторів, що використовують у якості міри подібності функцію відстаней. Результати оцінювання отримано на основі функції розподілу відстаней між об'єктами. При цьому розглянуто різні часткові випадки форми функції розподілу. Побудовано двосторонні верхні оцінки одиночного розпізнавання та розпізнавання для двох класів заданих розмірів. Запропоновано метод класифікації на основі співвідношення густин розподілу ймовірностей у нульовій та довільних точках.

6. Література

[1] Vorontsov, K.: On the influence of similarity of classifiers on the probability of overfitting pattern recognition and image analysis: new information technologies. In: Pattern Recognition and Image Analysis: new information technologies (PRIA-9). Volume 2., Nizhni Novgorod, Russian Federation (2008) 303–306

[2] Middleton, D.: The statistical theory of detection. i. optimum detection of signals in noise. In: Tech. Rep. Volume 35., Lincoln Lab., MIT (1953)

[3] Moon, T., Stirling, S.: Mathematical methods and algorithms for signal processing. Prentice/Hall, N.J. (2000)

[4] Wald, A.: Sequential Analysis. New York (1947)

[5] Vapnik, V.: The nature of statistical learning theory. 2 edn. Springer-Verlag, New York (2000)

[6] Kapustii, B., Rusyn, B., Tayanov, V.: Mathematical model of recognition systems with small databases. Journal of Automation and Information Sciences 39 (2007) 70–80

[7] Zhuravlev, J.: About the algebraic approach to recognition or classification tasks solution. Problems of cybernetics 33 (1978) 5–68

[8] Kapustii, B., Rusyn, B., Tayanov, V.: Estimation of the influence of information class coverage on generalized ability of the k-nearest-neighbors classifier. Automatic Control and Computer Sciences 42 (2008) 283–287

[9] Weisstein, E.: Chebyshev inequality from mathworld. A Wolfram Web Resource, http://mathworld.wolfram.com/ ChebyshevInequality.html (2010)