Оптимізація класифікатора в ознаковому і метричному просторах

Борис Капустій[1], Богдан Русин[2], Віталій Таянов[3]

[1]НУ "Львівська політехніка", Львів 79013, вул. С. Бандери 12, тел. (032)258-21-56

[2]ФМІ НАН України, Львів 79601, вул. Наукова 5а, т./ф. (032)229-61-09, rusyn@ipm.lviv.ua

[3]ФМІ НАН України, Львів 79601, вул. Наукова 5а, т./ф. (032)229-65-30, vtayanov@ipm.lviv.ua

Анотація. Наведені результати оптимізації класифікатора на основі перетворення Карунена-Лоєва (ПКЛ). В якості параметрів оптимізації вибрано вектори спектральних складових ПКЛ та геометричну міру відстаней Мінковського. Результатом оптимізації є мінімізація середньої кількості прототипів, необхідних для правильної ідентифікації тестованого образа із заданими ймовірностями помилок виявлення. Кількість прототипів визначається на основі процедур послідовного аналізу.

1. ВСТУП

При проектуванні системи розпізнавання образів (СР) важливо настроїти її так, щоб досягнути заданих значень імовірностей помилок виявлення 1-го та 2-го роду [6,8,9]. З цими значеннями однозначно пов'язана потрібна кількість прототипів класів. З іншого боку, технічні вимоги, що ставляться до СР, передбачають задання розміру класу або кількості прототипів, що відповідає класу. Тому якщо заданий розмір класу і помилки виявлення 1-го і 2-го роду, то задача зводиться до корекції класифікатора таким чином, щоб забезпечити вказані параметри. Будемо розглядати задачу розпізнавання з класами, що попарно не перетинаються. В загальному випадку класифікатор складається з деякого алгоритму розпізнавання, деякого вирішуючого правила про належність образа до одного з класів та з процедури визначення відстаней (або міри неподібності) в метричному або інформаційному просторі. Якщо зафіксувати середню кількість прототипів у межах класів, однозначно пов'язану з імовірностями помилок виявлення 1-го і 2-го роду, то регулюючими факторами будуть параметри самого алгоритму, вектор ознак (вектор спектральних складових ПКЛ), параметри вирішуючого правила та вибір метрики для проведення дискримінації класів у відповідному метричному просторі. При постійності всіх регулюючих параметрів класифікатора збільшення кількості прототипів призводить до зменшення помилок виявлення 1-го та 2-го роду за умови, що кількість класів залишиться незмінною.

Неможливість застосування класичних статистичних підходів в умовах малих вибірок обумовлюється тим, що середній розмір класу, необхідний для його ідентифікації із заданим рівнем вірогідності, має малу чутливість до зміни регулюючих параметрів класифікатора. Застосування ж запропонованого в [1-4] підходу щодо визначення вірогіднісних параметрів СР на основі інформації про окремо тестований образ дає можливість значно підвищити залежність середнього розміру класу, необхідного для його ідентифікації із заданим ступенем вірогідності, від зміни параметрів класифікатора. При цьому регулювання параметрів класифікатора можна здійснювати на тих вибірках, де класичні підходи вже не працюють.

2. ВИЗНАЧЕННЯ КРИТЕРІЇВ ПАРАМЕТРИЧНОЇ ОПТИМІЗАЦІЇ КЛАСИФІКАТОРА

Пропонуються наступні критерії для параметричної оптимізації класифікатора:

1. Пошук параметрів класифікатора в заданому діапазоні значень, при яких забезпечуються задана дисперсія і максимальне середнє значення показника оцінки вірогідності правильного розпізнавання для тестової послідовності образів:

$$\{\gamma_1, \gamma_2, ..., \gamma_n\} = \underset{\substack{\gamma_1 \in \Gamma_1, \\ \gamma_2 \in \Gamma_2, ..., \\ \gamma_n \in \Gamma_n}}{\operatorname{argmax}} E(I_{c_i} | \gamma_1, \gamma_2, ..., \gamma_n) | \operatorname{var}(\cdot) \le \operatorname{var}_0(\cdot), \quad (1)$$

де $\gamma_1, \gamma_2, ..., \gamma_n$ - параметри класифікатора; I_{c_i} - показник оцінки вірогідності правильного розпізнавання для i - го вхідного образу; $\operatorname{var}_0(\cdot)$ - задана дисперсія.

2. Пошук параметрів класифікатора в заданому діапазоні значень, при яких забезпечуються задане середнє значення і мінімальна дисперсія показника

оцінки вірогідності правильного розпізнавання для тестової послідовності образів:

$$\{\gamma_1, \gamma_2, ..., \gamma_n\} = \\ = \arg\min_{\substack{\gamma_1 \in \Gamma_1, \\ \gamma_2 \in \Gamma_2, ..., \\ \gamma_n \in \Gamma_n}} \mathrm{var}(I_{c_i} \mid \gamma_1, \gamma_2, ..., \gamma_n) \mid E(\cdot) \geq E_0(\cdot), \quad (2)$$

де $E_0(\cdot)$ - задане середнє.

Відтак, якщо в ролі показника оцінки вірогідності правильного розпізнавання окремо тестованого образа використовується коефіцієнт вірогідності розпізнавання K_{r_i} [1], то у вирази (1) та (2) замість I_{c_i} підставляється K_{r_i}. Якщо ж використати для оцінки вірогідності правильного розпізнавання ймовірність правильного розпізнавання $\hat{P}_{c_i}$, то замість I_{c_i} у вирази (1) та (2) слід підставляти $\hat{P}_{c_i}$ [4].

У випадку застосування для параметричного синтезу алгоритмів розпізнавання процедур послідовного аналізу (критерію Вальда) потрібно знайти області параметрів $\gamma_1 \in \Gamma_1, \gamma_2 \in \Gamma_2, ..., \gamma_n \in \Gamma_n$, які забезпечують мінімальний середній розмір класу.

3. ВИКОРИСТАННЯ ГЕОМЕТРИЧНИХ МІР ВІДСТАНЕЙ В ТЕОРІЇ РОЗПІЗНАВАННЯ ОБРАЗІВ

Основний перелік мір відстаней, що використовуються в теорії розпізнавання образів, наведений у [10]. Надалі буде застосовуватися міра відстаней Мінковського L_p, яка найчастіше використовується в теорії розпізнавання образів і є найбільш загальною мірою.

Якщо $p = 1$, то для міри відстаней Мінковського одержується лінійна міра відстаней L_1, а якщо $p = 2$ - Евклідова міра відстаней L_2. При $p = \infty$ одержується міра відстаней Чебишева, що еквівалентна максимальній різниці між відповідними елементами векторів. Це основні міри, що використовуються на практиці.

4. ЗАСТОСУВАННЯ ПОСЛІДОВНОГО АНАЛІЗУ ДЛЯ ДОСЛІДЖЕННЯ ВІРОГІДНОСТІ РОБОТИ КЛАСИФІКАТОРІВ

Застосовуючи послідовний аналіз [9] у зворотньому напрямку, можна визначити середні розміри класів s, необхідні для проведення розпізнавання з помилками виявлення 1-го роду (α) та 2-го роду (β) за умови, що θ є істинним параметром ($\theta = \{\theta_0, \theta_i\}, i = \overline{1, n\text{-}1}$). Якщо θ_0 відповідає правильному класу, а $\theta_i, i = \overline{1, n\text{-}1}$ – усім неправильним класам, то середні розміри відповідно дорівнюють [9]:

$$E_{\theta_0}(s) = \frac{I_0(\alpha,\beta)}{E_{\theta_0}(z)} = \frac{\beta \ln \frac{\beta}{1-\alpha} + (1-\beta)\ln(\frac{1-\beta}{\alpha})}{E_{\theta_0}(z)}; \\ E_{\theta_i}(s) = \frac{I_i(\alpha,\beta)}{E_{\theta_i}(z)} = \frac{(1-\alpha)\ln \frac{\beta}{1-\alpha} + \alpha \ln(\frac{1-\beta}{\alpha})}{E_{\theta_i}(z)}, \quad (3)$$

де $I_0(\alpha,\beta)$ та $I_i(\alpha,\beta)$ - функції, що визначають інформацію, необхідну для ідентифікації відповідно правильного та всіх неправильних класів із заданими помилками виявлення α та β; $E_\theta(z) = \int_0^\infty f(x \mid \theta) \ln \frac{f(x \mid \theta_0)}{f(x \mid \theta_i)} dx$ – інформація Кульбаха-Лейблера [10] для одного образу з правильного або неправильного класу; $f(x \mid \theta_0)$ та $f(x \mid \theta_i)$ – відповідні густини умовних імовірностей; $z_i = \ln \frac{f(x \mid \theta_0)}{f(x \mid \theta_i)}$.

На рис. 1 представлено функцію інформації $I(\alpha,\beta)$, що задовольняє умовам $I(\alpha,\beta) = I_i(\alpha,\beta)$ та $I(\alpha,\beta) = -I_0(\beta,\alpha)$. Для заданих значень α та β функція $I_0(\beta,\alpha)$ є константою. Якщо розрахований середній розмір правильного класу становить $s^0(\alpha^0,\beta^0)$, то розмір класу для довільних значень α та β розраховується як $s = s^0 \frac{I_0(\beta,\alpha)}{I_0(\beta^0,\alpha^0)}$. Таким чином, вірогідність роботи того чи іншого класифікатора цілком визначається параметрами $E_{\theta_0}(z)$ (3).

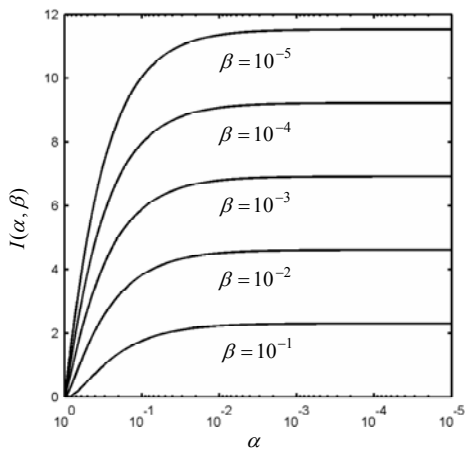

Рис.1. Функція інформації

Надалі будемо розглядати випадок нормально розподілених класів.

Нехай СР працює з базою даних, що налічує 40 класів облич людей по 18 реалізацій на кожний клас – усього 720 образів. У якості ознак використовуються спектральні компоненти ПКЛ.

На рис. 2 та 3 представлено залежності ймовірностей того, що середній розмір класу, необхідний для його інформаційного опису, заданого помилками виявлення α і β, перевищить фіксований розмір s_0 ($s_0 = 18$). Дані ймовірності розраховуються на основі нерівностей Чебишева, що відповідно задаються формулами:

$$P[|s - E(E_{\theta_0}(s^{(i)}))| \geq \varepsilon] \leq \frac{\mathrm{var}(E_{\theta_0}(s^{(i)}))}{\varepsilon^2}, \qquad (4)$$

$$\varepsilon = s_0 - E_{\theta_0}(s^{(i)}) > 0;$$

$$P[|s \geq s_0] \leq \frac{E_{\theta_0}(s^{(i)}))}{s_0}, (s_0 > 0, x \geq 0). \qquad (5)$$

Вираз (4) відповідає критерію параметричної оптимізації класифікатора шляхом мінімізації дисперсії параметрів оптимізації при збереженні прийнятного середнього значення, а вираз (5) – відповідно критерію, який передбачає максимізацію середнього значення при забезпеченні заданого рівня дисперсії параметрів оптимізації.

На рис.2 зображені залежності для значення p, що рівне 1.25 та 1.5 (рис.2), оскільки для них верхня межа ймовірності коректна. На рис. 3 практичний інтерес становлять криві з p, що дорівнює 0.25, 0.5, 0.75, 1, 1.25, 1.5 та 3. З точки зору регулярності характеристик і практичного використання доцільно брати метрики з $p = 1.25$ та 1.5.

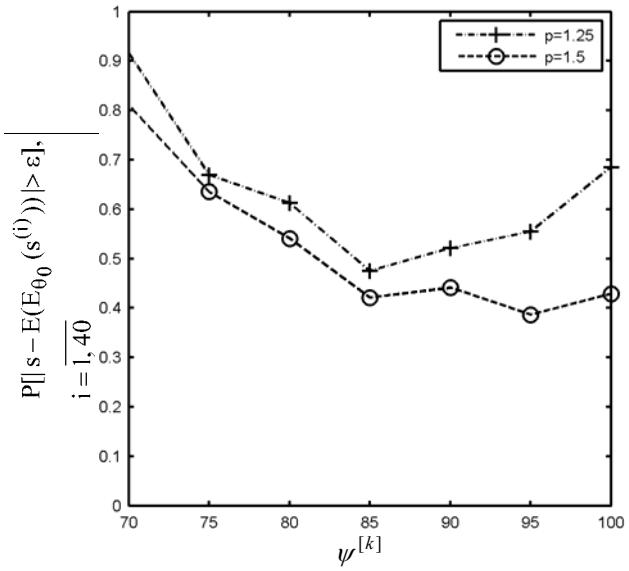

Рис.2. Залежність від кількості спектральних складових ПКЛ та показників p міри відстаней Мінковського ймовірності (4) перевищення потрібним середнім розміром класу його заданого вихідного розміру при $\alpha = 0.001$, $\beta = 0.1$

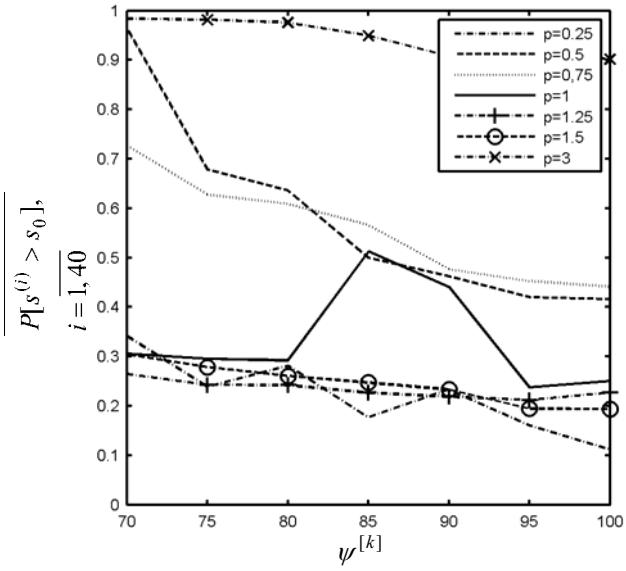

Рис.3. Залежність від кількості спектральних складових ПКЛ та показників p міри відстаней Мінковського ймовірності (5) перевищення потрібним середнім розміром класу його заданого вихідного розміру при $\alpha = 0.001$, $\beta = 0.1$

5. УСУНЕННЯ У ПРОЦЕСІ РОЗПІЗНАВАННЯ НАДЛИШКОВОЇ ІНФОРМАЦІЇ ЗОБРАЖЕНЬ ТА РОЗРАХУНОК ЕФЕКТУ ОПТИМІЗАЦІЇ

Відомо, що довільне зображення містить певну кількість надлишкової інформації з точки зору процедури розпізнавання. Оскільки цей надлишок не призводить до покращення вірогідності розпізнавання зображень, а часом навіть погіршує її, то намагаються вибрати найбільш оптимальну з точки зору вірогідності розпізнавання зображень та об'єму пам'яті СР, кількість ознак. З іншого боку, менша кількість ознак вимагає меншого часу на проведення порівняння зображень з точки зору тих чи інших критеріїв, а також часу на формування векторів ознак.

Визначимо виграш у кількості інформації, необхідної для збереження векторів ознак, а також у кількості операцій для розрахунку власних векторів розкладу ПКЛ та проведення процедури розпізнавання тестованого образу.

Оскільки зміна роздільної здатності зображення досить слабко впливає на низькочастотну частину спектра [4], то використання перших 100 коефіцієнтів розкладу при відновленні оригіналу зображення буде еквівалентним його роздільній здатності, що задається 100 дискретними відліками. Початкові зображення бази даних мають 644 дискретні відліки, а відношення висоти зображення до його ширини дорівнює 28/23. Якщо ж використовувати в зображенні 100 дискретних відліків, то відношення його сторін приблизно становитиме 11/9. З [5] відомо, що мінімальний середній розмір класу для заданих помилок

виявлення 1-го та 2-го роду дорівнює 5. У цьому випадку загальний розмір бази становить $5 \cdot 40 = 200$ зображень. Оскільки кожне зображення бази описується 100 коефіцієнтами розкладу ПКЛ, а тип даних short int у мові програмування C++ резервує 2 байта на коефіцієнт, то загальна кількість інформації для збереження бази даних дорівнюватиме $200 \cdot 100 \cdot 2 = 40000$ байт. Для збереження початкової бази даних потрібно $720 \cdot 644 \cdot 2 = 927360$ байт. Отже, виграш у кількості інформації для збереження векторів ознак зображень бази даних становить $927360 / 40000 \approx 23$ раза.

Визначимо також виграш у часових затратах на реалізацію ПКЛ та розпізнавання одного образа. Якщо розмір коваріаційної матриці дорівнює $n \times n$, то кількість операцій на розрахунок власних векторів та власних значень ПКЛ (операцій множення та додавання за методами Хаусхолдера та Гівенса) має порядок $O(n^3)$, а кількість інших операцій - порядок $O(n^2)$. Оскільки у випадку обробки зображень розмір матриці n, як правило, завжди великий, то відносний вплив операцій порядку $O(n^2)$ на часові затрати при реалізації ПКЛ буде незначним. Кількість операцій множення та додавання при обчисленні коваріаційної матриці також має порядок $O(n^2)$, тобто їх вплив теж малий. Оскільки вихідна коваріаційна матриця має розмір 644×644 елементи, а зменшена матриця, що отримується в результаті оптимізації класифікатора, - розмір 100×100 елементів, то виграш у кількості операцій в цьому випадку становить $(644/100)^3 \approx 267$ разів.

Насамкінець оцінимо виграш у кількості операцій, необхідних для проведення розпізнавання тестованого образа. Оскільки варіації ортогонального базису власних векторів здійснюються досить рідко, а саме у тих випадках, коли значно поповнюється база зображень, то в дійсності більш важливою стає оцінка виграшу в кількості операцій на етапі розпізнавання. Через те, що для забезпечення помилок виявлення 1-го та 2-го роду потрібно в середньому 5 образів на клас, виграш у кількості операцій складатиме $720 \cdot 644 / (200 \cdot 100) \approx 23$ раза.

6. РЕЗУЛЬТАТИ ПАРАМЕТРИЧНОГО СИНТЕЗУ СР

Алгоритм параметричного синтезу СР передбачає синтез такої системи з не гіршими від заданих значеннями показників вірогідності. В якості цих показників можна вибрати помилки виявлення 1-го та 2-го роду. Відтак, здійснюючи послідовний перебір наборів ознак та різних мір відстаней між зображеннями, проводять їх оптимізацію. В якості параметра оптимізації виступає середній розмір класу. Він визначається для оптимального значення набору ознак та певної міри відстаней між зображеннями. На цьому процедура параметричного синтезу завершується.

У випадку помилок виявлення 1-го та 2-го роду, рівних відповідно 0.001 та 0.1, параметричний синтез дає наступні результати: розмір вектора ознак становить 100 елементів; рекомендований порядок p метрики Мінковського складає 1.25 або 1.5; середній розмір класу рівний 5; розмір бази даних становить 200 зображень.

7. ВИСНОВКИ

Запропонований у роботі підхід до параметричної оптимізації класифікатора дає можливість отримати при реалізації алгоритму розпізнавання суттєвий виграш у часових та апаратних затратах порівняно з неоптимізованим класифікатором. При цьому вірогідність розпізнавання, що задовольняє заданим значенням помилок виявлення 1-го та 2-го роду, буде не гіршою, ніж до оптимізації. Подальший розвиток запропонованого підходу забезпечить можливість побудови класифікаторів, оптимальних за тими або іншими параметрами.

8. ЛІТЕРАТУРА

1. Капустій Б.О., Русин Б.П., Таянов В.А. Про особливості стохастичних характеристик коефіцієнта надійності розпізнавання для об'єктів з класу R^3. // Вісник НУ "Львівська політехніка". Радіоелектроніка та телекомунікації.-2003. – №477. – С 69-76.
2. Капустій Б.О., Русин Б.П., Таянов В.А. Розподіл середньоквадратичних відстаней між об'єктами в просторі R^2 // Відбір і обробка інформації. – 2003. – Випуск 19(95). – С.110-114.
3. Капустій Б.О., Русин Б.П., Таянов В.А. Новый подход к определению вероятности правильного распознавания объектов множеств // УСиМ. – 2005. – № 2. – С. 8-13.
4. Капустій Б.О., Русин Б.П., Таянов В.А. Нові підходи до параметричного синтезу алгоритмів розпізнавання // Радиоэлектроника и информатика. – 2005. – № 3. – С. 128-134.
5. Капустий Б.Е., Русын Б.П., Таянов В.А. Оптимизация классификаторов в условиях малых выборок // Автоматика и вычислительная техника. – 2005. – №3. – С. 72-86.
6. B.O. Kapustiy, B.P. Rusyn, V.A. Tayanov Peculiarities of Application of Statistical Detection Criteria for Problems of Pattern Recognition // Journal of Automation and Information Sciences. - 2005. - **37**, №2. - P. 30-36
7. E. Pekalska, R.P.W. Duin, S. Gunter, and H. Bunke, On not making dissimilarities Euclidean, in: A. Fred, T. Caelli, R.P.W. Duin, A. Campilho, and D. de Ridder (eds.), Structural, Syntactic, and Statistical Pattern Recognition, Proc. SSSPR2004 (Lisbon, Portugal, August 2004), Lecture Notes in Computer Science, vol. 3138, Springer Verlag, Berlin, 2004, 1145-1154.
8. Todd K. Moon, Wynn C. Stirling, Mathematical methods and algorithms for signal processing, Prentice/Hall, N.J., 2000.
9. A. Wald, Sequential Analysis, Wiley, New York, 1947.
10. A. R. Webb, Statistical Pattern Recognition, 2nd ed.. John Wiley & Sons, Inc., 2002.