УДК 004.93

В.А. ТАЯНОВ

ІМОВІРНІСНІ ТА СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ МОДЕЛІ ВИДІЛЕННЯ НАЙБІЛЬШ СТАБІЛЬНИХ ОБ'ЄКТІВ КЛАСІВ

Розглядається консенсусний підхід до прийняття рішень. Консенсус будується на основі двох алгоритмів, що є максимально несхожі за числом об'єктів, на яких не досягається консенсус. Підхід дає можливість виділяти найбільш стабільні об'єкти, на яких досягається консенсус максимально несхожих алгоритмів. Одночасно підхід дає можливість будувати оцінки ймовірностей попадання кожного об'єкта до однієї із трьох груп об'єктів: групи об'єктів, на яких досягається правильний консенсус, групи об'єктів, на яких консенсусні алгоритми одночасно помиляються, та групи об'єктів, на яких не досягається консенсус.

1. Вступ

Алгоритми прийняття рішень використовуються в таких задачах розпізнавання образів: розпізнавання з вчителем та розпізнавання без вчителя. Задачі розпізнавання без вчителя носять назву задач кластеризації і стосуються теорії кластерного аналізу. Задачі розпізнавання, де передбачено втручання оператора в процес розпізнавання, стосуються теорії навчання, зокрема машинного. Великий напрямок в теорії машинного навчання носить назву статистичного машинного навчання, започаткований в роботах В. Вапніка та Я. Червоненкіса в 60-70-х рр. минулого століття і продовжений у 90-х рр. того ж століття [1]. Задача кластерного аналізу формулюється так. Потрібно розбити вхідні дані на зони (кластери) згідно з прийнятим критерієм. Критерієм може виступати, наприклад, розкид даних відносно деякого центру, що називається центром кластера. В теорії кластеризації найбільш відомі два алгоритми: алгоритм внутрішньогрупових середніх (k-means algorithm) та алгоритм максимізації математичного очікування (Expectation maximization (EM) algorithm) [2]. Останній є найбільш популярним алгоритмом в останній час, однак має той недолік, що часто не знаходиться глобальний максимум. В задачах кластеризації часто вважається, що немає жодної апріорної інформації про належність даних до відповідних кластерів. Із задачами кластерного аналізу та із застосуванням різних алгоритмів і підходів можна ознайомитись в [3, 4].

Потрібно відзначити, що алгоритми побудови класифікаторів на основі навчаючих вибірок є нестабільними, оскільки нерегулярною є сама навчаюча вибірка. Внаслідок цього і виникла ідея у розробці інших алгоритмів, які частково використовують статистичне машинне навчання, однак мають значно меншу чутливість до неоднорідності навчаючих вибірок.

2. Формулювання задачі

В даній роботі увага приділяється задачам, що частково використовують навчання, тобто задачам машинного навчання. Згідно із загальною концепцією машинного навчання генеральна вибірка розбивається на навчаючу та тестову, або контрольну підвибірки. Для навчаючої підвибірки вважається, що належність об'єкта до свого класу є відомою. На контролі перевіряється надійність роботи алгоритму. Надійність роботи алгоритмів перевіряється методами ковзаючого контролю, яких є достатньо велика кількість різновидів [5,6].

Залежно від складності класифікації всі об'єкти можна розбити на три групи: об'єкти, які є стабільними і класифікуються з великою надійністю, об'єкти, що належать пограничній зоні між класами, та об'єкти, що належать одному класу та занурені глибоко в середину іншого ласу [7]. Серед тих об'єктів, які можуть спричинити помилку, найбільшу частину складають пограничні об'єкти. Тому важливим є розробити алгоритм, що дозволяє виділяти найбільшу кількість пограничних об'єктів. Оскільки класифікація пограничних об'єктів є ненадійною. то потрібно застосувати спеціальні верифікаційні алгоритми щодо визначення класів, до яких дійсно належать пограничні об'єкти. Однак дана робота присвячена першо-

му етапу ієрархічного алгоритму класифікації – розробці консенсусного алгоритму для детектування пограничних об'єктів. Також задача полягає у визначенні ймовірностей попадання об'єкта у кожну із згаданих вище груп об'єктів.

Оскільки математичний апарат, що використовується в даній роботі, відноситься до побудови комбінованих або ієрархічних алгоритмів, то важливо присвятити декілька розділів розгляду задачі класифікації саме з цієї точки зору.

3. Комбінування результатів класифікації

Розглянемо задачу класифікації об'єктів на n класів. Вважаємо, що класифікація здійснюється ансамблем або композицією класифікаторів. Нехай I – число об'єктів, а P – число алгоритмів класифікації. Кожний алгоритм (позначається індексом p, $p = 1,...p$) асоціює кожний об'єкт з одним і лише одним класом. Можна зобразити результат роботи p-го алгоритму за допомогою бінарної прямокутної матриці B^p з I рядками і J_p стовпцями, де J_p – це число класів:

$$B_i^k = \begin{cases} 1, & i \in k; \\ 0, & i \notin k, \end{cases}$$

тут $i = 1,...I$, $k = 1,...,J_p$, а B^p називається блочною матрицею. Перелічимо властивості матриці B^p:

1) Всі колонки матриці B ортогональні.
2) $\sum_{k=1}^{J_p} B_{ik}^p = 1$.
3) Якщо об'єкти i та l з одного класу (мають однакову мітку), то $\sum_{k=1}^{J_p} B_{ik}^p B_{il}^p = 1$, інакше $\sum_{k=1}^{J_p} B_{ik}^p B_{il}^p = 0$.

Представимо матрицю B у вигляді об'єднання матриць B^p так, що

$$B = [B^1,...,B^p,...,B^p].$$

Тут відстань між об'єктами i та l може бути представлена у вигляді добутку B_i та B_l. Якщо об'єкти i та l завжди в тому самому класі для різних алгоритмів P, то скалярний добуток векторів $B_i B_l'$ рівний P, в іншому випадку $B_i B_l' = 0$. Скалярний добуток

$$B_i B_l' = \sum_{p=1}^{P} \sum_{k=1}^{J_p} B_{ik} B_{lk} = 0$$

має дві важливі властивості:

1) обмеження знизу значенням нуля;
2) обмеження значенням P зверху.

Після нормалізації на P отримаємо $0 \le B_i B_l' \le 1$.

Відстань між об'єктами i та l може бути представлена у вигляді:

$$d(B_i B_l') = 1 - \sum_{p=1}^{P} \sum_{k=1}^{J_p} B_{ik} B_{lk}.$$

4. Блочні алгоритми

Визначимо евклідову відстань $d_E(B_i B_l)$ між точками B_i та B_l:

$$d_E(B_i, B_l) = (B_i - B_l)^2 = \sum_{p=1}^{P} \sum_{k=1}^{J_p} (B_{ik} - B_{lk})^2 = 2P - 2B_i B_l'.$$

Хемінгова відстань між об'єктами i та l представляється у вигляді:

$$d_H(B_i, B_l) = \frac{1}{2} d_E(B_i, B_l) = P - B_i B_l'.$$

Хемінгова відстань показує число біт, на які відрізняються дві двійкові послідовності.

5. Імовірнісна модель кластеризації на основі розподілу Бернуллі

Оскільки матриця B є бінарною, то вважаємо, що всі її колонки мають розподіл Бернуллі з параметром μ_i:

$$P(B \mid \mu) = \prod_{j=1}^{J} \mu_j^{B_j} (1-\mu_j)^{1-B_j},$$

де $\mu = \{\mu_1, ..., \mu_j, ..., \mu_J\}$. Середнє значення розподілу рівне $E[B] = \mu$, а коваріаційна матриця дорівнює $\mathrm{cov}[B] = \mathrm{diag}\{\mu_j(1-\mu_j)\}$.

Суміш з k компонент розподілу Бернуллі для об'єкта B_i має вигляд

$$P(B_i \mid \mu, \alpha) = \sum_{k=1}^{K} \alpha_k P_k(B_i \mid \mu_k),$$

де $\mu = \{\mu_{1j}, ..., \mu_{kj}, ..., \mu_{Kj}\}$, а ймовірність для B_i даного μ_k є

$$P_k(B_i \mid \mu_k) = \prod_{j=1}^{J} \mu_{kj}^{B_{ij}} (1-\mu_{kj})^{1-B_{ij}}.$$

Знайдемо правдоподібність для об'єкта B_i, даних μ і α, якщо дані у B незалежні та ідентично розподілені:

$$\log P(B \mid \mu, a) = \log \prod_{i=1}^{I} P(B_i \mid \mu, \alpha) = \sum_{i=1}^{I} \log \left\{ \sum_{k=1}^{K} \alpha_k P_k(B_i \mid \mu_k) \right\}.$$

Використовуючи теорему Баєса, можна отримати ваги w_{ik} або умовну ймовірність того, що значення B_i належить класу k:

$$w_{ik} = \frac{\alpha_k P_k(B_i \mid \mu_k)}{\sum_{l=1}^{K} \alpha_l P_l(B_i \mid \mu_l)}.$$

Для визначення параметрів розподілу Бернуллі пропонується EM – алгоритм. Параметри μ та α обчислюються на кроці M, а ваги обчислюються на кроці E. Очікувана кількість об'єктів N_k компонента k є: $N_k = \sum_{i=1}^{I} w_{ik}$.

Середнє значення μ_k компонента дорівнює

$$\mu_{kj} = \frac{1}{N_k} \sum_{i=1}^{I} w_{ik} B_{ij} \text{ і } \alpha_k = \frac{N_k}{N}.$$

Відзначимо, що мають виконуватись такі нерівності:

$$\sum_{k=1}^{K} \alpha_k = 1;$$
$$\sum_{j=1}^{J} \mu_{kj} = 1;$$
$$0 \le \alpha_k \le 1;$$
$$0 \le \mu_k \le 1.$$

На рис.1 наведені псевдокоди EM – алгоритму для розподілу Бернуллі.

1: Initialise K means μ_k and α_k.

2: **E-step**

2.1: Calculate probabilities $P_k(B_i \mid \mu_k)$

$$P_k(B_i \mid \mu_k) = \prod_{j=1}^{J} \mu_{kj}^{B_{ij}} (1 - \mu_{kj})^{(1-B_{ij})}$$

2.2: Calculate weights w_{ik}

$$w_{ik} = \frac{\alpha_k P_k(B_i \mid \mu_k)}{\sum_{l=1}^{K} \alpha_l P_l(B_i \mid \mu_l)}$$

4: **M-step** Re-estimate parameters N_k

4.1: $$N_k = \sum_{i=1}^{I} w_{ik}.$$

4.2: $$\mu_{kj} = \frac{1}{N_k} \sum_{i=1}^{I} w_{ik} B_{ij}.$$

4.3: $$\alpha_k = \frac{N_k}{N}.$$

5: Evaluate the log-likelihood function:

$$\log P(B \mid \mu, \alpha) = \sum_{i=1}^{I} \log \left\{ \sum_{k=1}^{K} \alpha_k P_k(B_i \mid \mu_k) \right\}$$

6: If log-likelihood is converged,

then stop,

else go to Step 2.

Рис.1. Псевдокод EM – алгоритму для розподілу Бернуллі

6. Поліноміальна модель

Поліноміальний розподіл застосовується тоді, коли елементи бінарної матриці є взаємовиключними. Якщо p – число різних кластеризацій, кожна з яких має K^p кластерів, $p = 1,...,P$, $j_p = 1,...,K^p$ – індекс кластеру j_p в кластеризації P. Ймовірність того, що $B_{ij_p} = 1$, рівна μ_{ij_p}. Тоді

$$P(B_{ij_p}) = \prod_{j_p=1}^{J_p} \mu_{ij_p},$$

тут $\mu_{ij_p} > 0$ і $\sum_{j=1}^{J_p} = 1, \forall i, p$. Поліноміальний розподіл при цьому запишеться у вигляді:

$$P(B \mid \mu) = \prod_{p=1}^{P} \prod_{j_p=1}^{J_p} \mu_{j_p}^{B_{j_p}},$$

де $\mu = \{\mu_1,...,\mu_{j_p},...,\mu_{J_p}\}$. Суміш з k компонент для поліноміального розподілу запишеться у вигляді:

$$P(B_i \mid \mu, \alpha) = \sum_{k=1}^{K} \alpha_k P_k(B_i \mid \mu_k).$$

Функція відношення правдоподібностей та умовна ймовірність того, що $B_i \in k$, представляється аналогічно біноміальному випадку.

Очікувана кількість елементів N_k k-го компонента та інші елементи, що входять у модель суміші, представляється так:

$$N_k = \sum_{i=1}^{I} w_{ik};$$

$$\alpha_k = \frac{N_k}{N};$$

$$\mu_{kj_p} = \frac{1}{N_k} \sum_{i=1}^{I} B_{ij_p}.$$

Для ініціалізації коефіцієнтів α_k часто приймають, що $\alpha_k = 1/K$.

На рис.2 наведені псевдокоди ЕМ – алгоритму для поліноміального розподілу.

1: Initialise K means μ_k and α_k.

2: **E-step**

2.1: Calculate probabilities $P_k(B_i \mid \mu_k)$

$$P_k(B_i \mid \mu_k) = \prod_{p=1}^{P} \prod_{j_p=1}^{J_p} \mu_{kj_p}^{B_{ijp}}$$

2.2: Calculate weights w_{ik}

$$w_{ik} = \frac{\alpha_k P_k(B_i \mid \mu_k)}{\sum_{l=1}^{K} \alpha_l P_l(B_i \mid \mu_l)}$$

4: **M-step** Re-estimate parameters N_k

4.1: $$N_k = \sum_{i=1}^{I} w_{ik}.$$

4.2: $$\mu_{kj_p} = \frac{1}{N_k} \sum_{i=1}^{I} B_{ij_p}.$$

4.3: $$\alpha_k = \frac{N_k}{N}.$$

5: Evaluate the log-likelihood function:

$$\log P(B \mid \mu, \alpha) = \sum_{i=1}^{I} \log \left\{ \sum_{k=1}^{K} \alpha_k P_k(B_i \mid \mu_k) \right\}$$

6: If log-likelihood is converged,

then stop,

else go to **Step 2**.

Рис.2. Псевдокод ЕМ – алгоритму для поліноміального розподілу

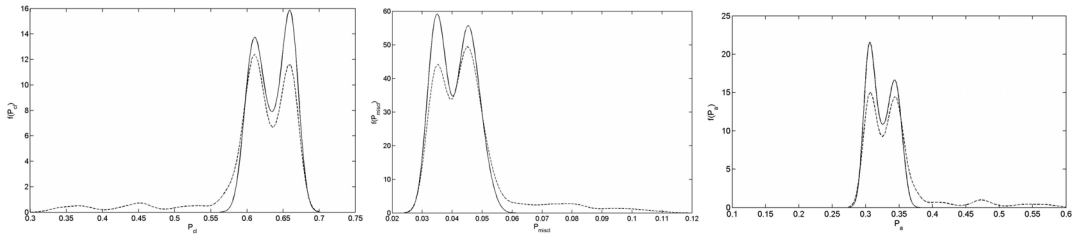
7. Модель кластеризації на основі визначення найбільш стабільних об'єктів

Ідея побудови даної моделі полягає в тому, що загальна сукупність об'єктів, які підлягають класифікації розбивається на три функціональні групи. До першої групи належать об'єкти, що мають високу надійність класифікації. Висока надійність класифікації означає, що об'єкти класифікуються коректно при сильних (максимальних) відхиленнях параметрів від оптимальних. З точки зору складності класифікації ці об'єкти відносяться до групи легких об'єктів. До другої групи належать об'єкти, на яких немає консенсусу. Якщо вибрати два алгоритми у композиції алгоритмів, то вони повинні бути максимально несхожі [8] і на них не повинно бути консенсусу. Якщо використовувати більшу кількість алгоритмів, то об'єкт належатиме до другої групи, якщо немає консенсусу на всіх алгоритмах. Якщо для побудови консенсусу використовувати проміжні алгоритми, параметри яких знаходяться в межах інтервалів між параметрами двох найбільш неподібних алгоритмів, то це не дає можливості виділяти більшу кількість об'єктів, на яких немає консенсусу. Неподібність між алгоритмами визначається на основі хемінгової відстані між результатами роботи двох алгоритмів, заданих у вигляді двійкових послідовностей. На практиці це означає також, що в цілому не будуть виділятись і нові за складом об'єкти, якщо використовувати консенсус із

більшого числа алгоритмів. Третя група складається з тих об'єктів, на яких помиляються обидва алгоритми і при цьому вони знаходяться у консенсусі. Помилка, що обумовлюється цими об'єктами, не може бути зменшена вже за жодних умов. Таким чином, помилка не може бути меншою за значення, що обумовлюється відносною долею об'єктів із третьої групи. Наступним кроком буде перекласифікація об'єктів із другої групи, тобто визначення, до якого саме класу належить той чи інший об'єкт.

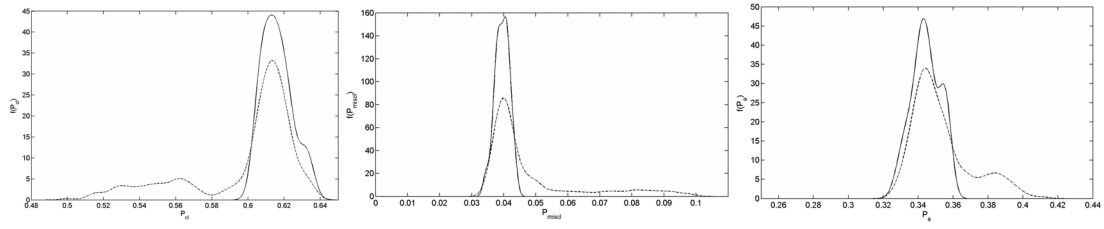
Дослідження, що проводяться в цій роботі, стосуються аналізу статистичних характеристик результатів консенсусу, побудованих на основі двох алгоритмів. Задачею аналізу є визначення регулярності статистичних характеристик на різних підвибірках, взятих шляхом розбиття генеральної вибірки на блоки різних розмірів. Розподіли ймовірностей по консенсусу для трьох груп об'єктів здійснювались непараметричним оцінюванням за допомогою вікна Парзена з використанням гаусівських ядер.

На рис. 3-6 наведені графічні залежності результатів консенсусу для задач, взятих із репозиторію UCI. Цей репозиторій сформований у Каліфорнійському університеті. Структура даних в задачах із цього репозиторію є такою. Кожна задача записана у вигляді текстового файлу, де стовпцями є ознаки того чи іншого об'єкта, а рядки складаються із сукупності ознак для того чи іншого об'єкта, тобто кількість рядків відповідає кількості об'єктів, а кількості стовпців – число ознак для кожного об'єкта. Окремий стовпець складається із міток класів, якими позначений кожний об'єкт. Дуже багато задач із цього репозиторію стосуються таких галузей, як біологія та медицина. На рис. 3-6 суцільною лінією позначено розподіл результатів, отриманих за допомогою алгоритмів крос-валідації на підвибірках з мінімальним розміром Q=200, а штрихпунктирною – на підвибірках із мінімальним розміром Q=30.



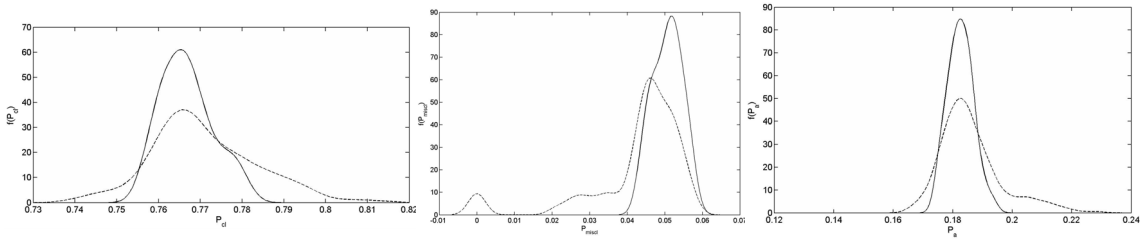
а б в

Рис. 3. Задача " pima-indians-diabetes" із репозиторію UCI: а – непараметрично оцінена густина розподілу ймовірностей правильного консенсусу із двох алгоритмів;б – непараметрично оцінена густина розподілу ймовірностей неправильного консенсусу із двох алгоритмів; в – непараметрично оцінена густина розподілу ймовірностей неконсенсусу із двох алгоритмів



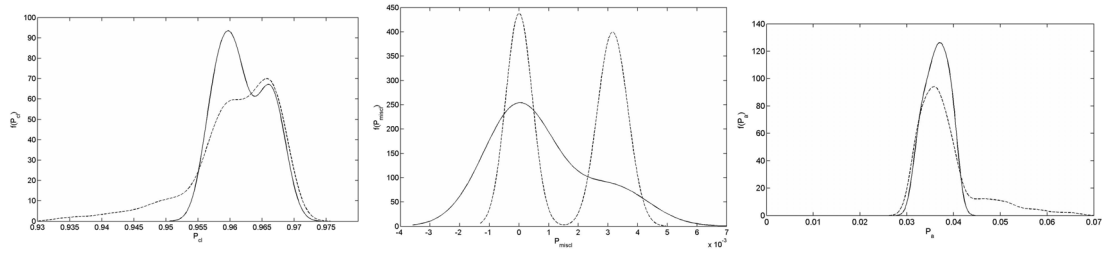
а б в

Рис. 4. Задача "bupa" із репозиторію UCI: а – непараметрично оцінена густина розподілу ймовірностей правильного консенсусу із двох алгоритмів;б – непараметрично оцінена густина розподілу ймовірностей неправильного консенсусу із двох алгоритмів;в – непараметрично оцінена густина розподілу ймовірностей неконсенсусу із двох алгоритмів



а б в

Рис. 5. Задача “haberman” із репозиторію UCI: а – непараметрично оцінена густина розподілу ймовірностей правильного консенсусу із двох алгоритмів;б – непараметрично оцінена густина розподілу ймовірностей неправильного консенсусу із двох алгоритмів;в – непараметрично оцінена густина розподілу ймовірностей неконсенсусу із двох алгоритмів



а б в

Рис. 6. Задача “dermatology ” із репозиторію UCI: а – непараметрично оцінена густина розподілу ймовірностей правильного консенсусу із двох алгоритмів; б – непараметрично оцінена густина розподілу ймовірностей неправильного консенсусу із двох алгоритмів; в – непараметрично оцінена густина розподілу ймовірностей неконсенсусу із двох алгоритмів

У табл. 1-4 наведені оцінки ймовірностей попадання кожного об’єкта із задач репозиторію UCI у кожну із трьох груп об’єктів. У даному випадку об’єкти, на яких існує консенсус найбільш неподібних алгоритмів, відносяться до “простих” об’єктів, а об’єкти, на яких обидва алгоритми, що утворюють помилковий консенсус, відносяться до групи об’єктів, що спричиняють некоректовану помилку, яку не можна зменшити в принципі; об’єкти, на яких немає консенсусу, відносяться до групи пограничних об’єктів. У табл.1-4 також наведені дисперсії відповідних ймовірностей. Мінімальний розмір блоків, на основі яких побудовані оцінки, в одному випадку складає 30 об’єктів, а в іншому – 200.

Таблиця 1. Задача “pima” із репозиторію UCI

	Q=200		Q=30	
	μ	σ	μ	σ
Ймовірність консенсусу	0.635	0.024	0.611	0.064
Помилка консенсусу	0.041	0.006	0.046	0.013
Ймовірність неконсенсусу	0.324	0.019	0.344	0.052

Таблиця 2. Задача “bupa” із репозиторію UCI

	Q=200		Q=30	
	μ	σ	μ	σ
Ймовірність консенсусу	0.616	0.008	0.599	0.030
Помилка консенсусу	0.040	0.002	0.048	0.016
Ймовірність неконсенсусу	0.344	0.008	0.353	0.017

Таблиця 3. Задача “haberman” із репозиторію UCI

	Q=200		Q=30	
	μ	σ	μ	σ
Ймовірність консенсусу	0.767	0.006	0.771	0.013
Помилка консенсусу	0.051	0.004	0.043	0.013
Ймовірність неконсенсусу	0.183	0.004	0.186	0.011

Таблиця 4. Задача “dermatology” із репозиторію UCI

	Q=200		Q=30	
	μ	σ	μ	σ
Ймовірність консенсусу	0.962	0.004	0.961	0.007
Помилка консенсусу	0.002	0.002	0.001	0.001
Ймовірність неконсенсусу	0.036	0.003	0.038	0.007

У табл. 5 наведена доля помилок на тестових даних при тестуванні різних алгоритмів класифікації для двох задач із репозиторію UCI. Для запропонованого методу наведені мінімальні та максимальні помилки, що можна отримати на приведених тестових задачах.

Таблиця 5. Порівняння результатів класифікації різними методами

Метод / Задача	bupa	pima
Monotone (SVM)	0.313	0.236
Monotone (Parzen)	0.327	0.302
AdaBoost (SVM)	0.307	0.227
AdaBoost (Parzen)	0.33	0.290
SVM	0.422	0.230
Parzen	0.338	0.307
RVM	0.333	-
Запропонований метод (min/max)	0.040/0.212	0.041/0.203

В табл. 5 для запропонованого алгоритму значення мінімальної помилки дорівнює помилці консенсусу, а максимальної – сумі мінімальної та половині помилки неконсенсусу. Як видно із табл. 5, значення максимальної помилки значно менше від найменшого значення помилки всіх наведених алгоритмів для обох задач із репозиторію UCI. В порівнянні із деякими алгоритмами значення мінімальної помилки для запропонованого алгоритму менше на порядок. Запропонований метод характеризується значно більшою стабільністю помилок класифікації за інші методи.

8. Аналіз отриманих статистичних характеристик

На рис. 3-6 наведені параметрично оцінені густини розподілів ймовірностей для ймовірності правильного консенсусу, ймовірності неправильного консенсусу та ймовірності неконсенсусу. Як можна побачити з цих рисунків, дані розподіли можуть бути представлені за допомогою однокомпонентної, двокомпонентної або багатокомпонентної моделі розподілів. Багатокомпонентна модель задається сумішшю гаусіан, які входять із своїми коефіцієнтами впливу. Параметри розподілів та коефіцієнтів участі у моделі оцінюються за допомогою ЕМ – алгоритму. Оцінювання значень відповідних ймовірностей здійснювалось блоками з мінімальним розміром у $Q = 30$ та $Q = 200$ елементів. Дані розміри блоків обумовлюються розмірами малих вибірок, які за різними критеріями коливаються в межах від 30 до 200 елементів [9]. Згідно із стандартним означенням малою вибіркою називається вибірка, що характеризується нерегулярними статистичними характеристиками. Як видно із всіх рисунків, оцінки по блоках із мінімальним розміром у 30 елементів є нерегулярними, що вказує на те, що для даних задач підвибірки розміром у 30 елементів і дещо більшим є малими. На це вказують довгі хвости у відповідних розподілах ймовірностей. Максимум у нульовій точці для двокомпонентних моделей характеризується великою кількістю нульових ймовір-

ностей. Це може бути, коли немає помилок у роботі консенсусу із двох алгоритмів. Оцінки відповідних ймовірностей на основі середніх значень та за максимумом відповідних розподілів імовірностей (оцінка за максимумом правдоподібності) не сильно відрізняються, що дає додаткову гарантію на правдоподібність відповідних оцінок. Значення отриманих оцінок ймовірностей правильного консенсусу, неправильного консенсусу та ймовірності неконсенсусу дає можливість оцінити складність задач з точки зору класифікації. Задачі та алгоритми оцінки складності задач розпізнавання розглянуті у [10]. Так, задачі "pima" і "bupa" є приблизно однакові, оскільки значення зазначених трьох ймовірностей є приблизно однаковими. Задача "haberman" є менш складною, а задача "dermatology" є найпростішою із всіх чотирьох задач. Математичний аналіз побудови композицій алгоритмів детально розглянутий у [11].

9. Висновки

В роботі визначені ймовірності попадання в кожну із трьох груп об'єктів: група простих об'єктів, на яких досягається правильний консенсус із двох алгоритмів, група об'єктів, на яких досягнутий неправильний консенсус із двох алгоритмів, та група об'єктів на яких не досягається консенсус. Аналіз показує, що існують розподіли даних ймовірностей, які можна представляти у вигляді багатокомпонентних моделей, зокрема багатокомпонентної моделі суміші гаусіан. Все це дає можливість аналізувати запропоновані алгоритми методами математичної статистики та теорії ймовірностей. З наведених рисунків і таблиць видно, що оцінки ймовірностей при використанні методів ковзаючого контролю з усередненням по блоках з мінімальним розміром у 30 та 200 елементів мало відрізняються між собою, що дає можливість зробити висновок про те, що такий метод побудови консенсусу, де у консенсусі беруть участь найбільш несхожі алгоритми, є достатньо регулярним і не має такої чутливості до вибірки, як інші алгоритми, що використовують навчання. Як видно із відповідних таблиць, мінімальна помилка класифікації є практично на порядок меншою від найкращих існуючих алгоритмів, а максимальна помилка – меншою від 1,5 до 2-х разів. Також відповідні помилки є значно стабільніші як відносно задачі, на якій тестується метод, так і відносно серії наведених алгоритмів, де значення помилки має достатньо велику дисперсію. Більше того, оскільки значення мінімальної помилки є достатньо малим і стабільним, то це гарантує стабільність отримання коректних результатів класифікації на об'єктах, на яких досягається консенсус максимально несхожих алгоритмів. Відносно інших алгоритмів такої впевненості не буде. Дійсно, значення помилки на рівні 30-40% (у порівнянні із 4%) не дає жодної впевненості у результатах класифікації.

Список літератури: 1. *Vapnik V.* The nature of statistical learning theory. New York: Springer-Verlag, 2 edn, 2000. **2.** *Bishop C.* Pattern recognition and machine learning. New York: Springer, 2006. **3.** *Kyrgyzov I.* Recherche dans les bases de donnes satellitaires despaysages et application au milieu urban :clustering, consensus et catgorisation. Ph.D. thesis. Paris: L'cole Nationale Supriere des Tlcommunications, 2008. **4.** *Taylor, J., Cristianini, N.*: Kernel methods for pattern analysis. New York: Cambridge University Press, 2004. **5.** *Kohavi, R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // 14th International Joint Conference on Artificial Intelligence, Palais de Congres, Montreal, Quebec, Canada, 1995. P. 1137–1145. **6.** *Mullin, M., Sukthankar, R.* Complete cross-validation for nearest neighbor classifiers // Proceedings of International Conference on Machine Learning, 2000. P. 639–646. **7.** *Vorontsov K.* Combinatorial approach to quality estimation of learning algorithms // Mathematical questions of cybernetic, 13, 2004. P. 5–36. **8.** *Vorontsov K.* On the influence of similarity of classifiers on the probability of overfitting pattern recognition and image analysis: new information technologies // Pattern Recognition and Image Analysis: new information technologies (PRIA-9), Volume 2, Nizhni Novgorod, Russian Federation, 2008. P. 303–306. **9.** *Гуров С.И.* Оценка надёжности классифицирующих алгоритмов. М.: Издательский отдел ф-та ВМиК МГУ, 2003. 45 с. **10.** *Basu M.* Data complexity in pattern recognition. London: Springer, 2006. **11.** *Zhuravlev, J.* About the algebraic approach to recognition or classification tasks solution // Problems of cybernetics, 33, 1978. P. 5–68.

Таянов Віталій Анатолійович, канд. техн. наук, старший викладач ЛДІНТУ ім. В'ячеслава Чорновола. Наукові інтереси: математичні методи розпізнавання образів. Адреса: Україна, 79057, Львів, вул. генерала Чупринки, 130, тел. 237-80-73, e-mail: vtayanov@yahoo.com.