УДК 004.9

Валентина Миколаївна Панченко

ВИБІР ПОРОГОВОГО ЗНАЧЕННЯ ЙМОВІРНІСНОГО КЛАСИФІКАТОРА ТЕКСТІВ

Постановка задачі. Інтерес до вирішення задачі автоматизованої класифікації текстів за наперед заданими категоріями викликаний впровадженням електронного документообігу та потребою в упорядкуванні документів, що мають електронну форму. Адже своєчасне отримання повної та точної інформації є запорукою інформаційної безпеки будь-якої організації.

У загальному вигляді задача класифікації текстів може бути сформульована таким чином [1].

Маємо множину категорій (класів, ознак) $C = \{c_1, \ldots, c_{|C|}\}$.

Маємо множину текстових документів $D = \{d_1, \ldots, d_{|D|}\}$.

Невідома цільова функція $\Phi : C \times D \to \{0, 1\}$.

Необхідно побудувати класифікатор Φ', максимально близький до Φ. Класифікатор може надавати точну відповідь $\Phi' : C \times D \to \{0, 1\}$ або ступінь подібності $\Phi' : C \times D \to [0, 1]$.

Серед відомих підходів до вирішення поставленої задачі найбільшої популярності на сьогодні набули алгоритми класифікації за навчальною вибіркою [2]. Вони забезпечують побудову класифікуючого правила, т. зв. класифікатора, в процесі індуктивної автоматизованої обробки попередньо згрупованої за необхідними категоріями множини документів $R \subset C \times D$, для яких значення Φ відомі. Такий підхід отримав назву "машинне навчання" (англ. *machine learning*). Його перевагами у порівнянні з підходом інженерії знань (англ. *knowledge engineering*), що передбачає побудову класифікуючого правила вручну експертами предметної галузі, є висока ефективність, скорочення часових витрат та інтелектуальних зусиль експертів, можливість застосування у різних прикладних сферах, зокрема для фільтрації спаму, тематичної рубрикації, визначення емоційного забарвлення, розподілу контекстної реклами.

Вирішення задачі автоматизованої класифікації текстів із використанням машинного навчання структурно складається з таких етапів: вибір моделі представлення текстових даних, навчання (побудова класифікуючого правила за даними навчальної вибірки), оцінювання документів. Так, для представлення текстових даних може бути використано моделі, запозичені з теорії інформаційного пошуку: векторно-просторова, булева, ймовірнісна [3]. Для побудови класифікуючого правила використовують методи: опорних векторів, дерева рішень, нейронні мережі, наївний байесівський підхід тощо [1].

У статті розглядатимемо класифікатор, який використовує ймовірнісну модель представлення текстових даних і ґрунтується на аксіоматичному припущенні про взаємну незалежність входження термів (морфемних основ слів) до тексту, у зв'язку з чим цей підхід отримав назву "наївний байесівський" (англ. *naive Bayesian*). Основним його принципом є ймовірнісна оцінка ваги терму в документі. З іншого боку, оцінювання відповідності документа, що аналізується, певній категорії здійснюється на основі ймовірності того, що експерт визнав належність документів навчальної вибірки до певної категорії.

Математична модель ймовірнісного класифікатора може бути подане таким чином. На *етапі навчання* маємо множину документів $D = \{d_1, \ldots, d_{|D|}\}$, *згруповану за необхідними категоріями* $C = \{c_1, \ldots, c_{|C|}\}$, *та словник усіх різних термів* $T = \{t_1, \ldots, t_{|T|}\}$, що входять у документи даної навчальної вибірки. Позначимо через A_j подію, яка полягає в наявності терму t_j в тексті документа d, що аналізується, $j = 1..|T|$; через C_i — подію, яка полягає в тому, що документ d відноситься до категорії i, $i = 1..|C|$. Зауважимо, що події C_i становлять повну групу подій. Тоді $P(A_j|C_i)$ — ймовірність зустріти терм t_j у документі d, що відноситься до категорії i. Відповідно, умовна ймовірність того, що документ d відноситься до категорії i, якщо відомо, що він містить терм t_j, визначається за формулою Байеса:

$$P(C_i | A_j) = \frac{P(A_j | C_i) P(C_i)}{\sum_{i=1}^{|C|} P(A_j | C_i) P(C_i)}.$$

Назвемо цю величину *ідентифікаційною потужністю* терму t_j і позначимо Ip_j.

Для оцінювання ймовірностей $P(A_j|C_i)$, $P(C_i)$ використовують навчальну вибірку документів, наприклад:

$$\hat{P}(A_j|C_i)=\frac{k^j_{C_i}}{n_{C_i}};\qquad \hat{P}(C_i)=\frac{m_{C_i}}{m},$$

де $k^j_{C_i}$ — кількість термів t_j у групі документів категорії i;

n_{C_i} — загальна кількість термів у групі документів категорії i;

m_{C_i} — кількість документів категорії i;

m — загальна кількість документів навчальної вибірки, $m=\sum_{i=1}^{|C|} m_{C_i}$.

Отримані оцінки ідентифікаційної потужності термів, зважені на множині документів навчальної вибірки, є базовою інформацією для оцінювання ймовірності належності документа d, що аналізується, до категорії i за формулою [4]:

$$\hat{P}(C_i|\bigcap_{j:t_j\in d}A_j)=\frac{\prod_{j:t_j\in d} Ip_j}{\prod_{j:t_j\in d} Ip_j+\prod_{j:t_j\in d}(1-Ip_j)}.$$

Така оцінка отримує свої значення в інтервалі [0, 1], а тому її фізичний зміст може бути проінтерпретований як ймовірність, з якою досліджуваний документ належить до певної категорії. Отже, результатом роботи ймовірнісного класифікатора є ступінь подібності документа за контентом до документів певної категорії: $\Phi': C\times D\to[0,1]$.

Водночас, на практиці часто вирішуються завдання класифікації множини документів за двома категоріями, тобто коли $|C|=2$: спам / не спам, позитивний / негативний, маніпулятивний / інформативний текст тощо. У такому випадку більш корисною є точна відповідь класифікатора $\Phi': C\times D\to\{0,1\}$. Перехід від шкали подібності до двоелементної шкали (належить / не належить до певної категорії) пропонується здійснювати шляхом порівняння отриманої оцінки ймовірності з деяким пороговим значенням Z, яке обирається на основі досвіду та інтуїції дослідника [3; 4]. На думку автора, такий підхід до вибору поргового значення ставить результати автоматизованої класифікації в залежність від суб'єктивних факторів. Тому ***метою статті*** є побудова формалізованої процедури для вибору порогового значення ймовірнісного класифікатора у випадку двоелементної множини категорій. Відповідно необхідно вирішити такі ***завдання:*** побудувати алгоритм ймовірнісного класифікатора у випадку двоелементної множини категорій; обґрунтувати процедуру вибору порогового значення; визначити напрями подальших досліджень, пов'язані з вирішенням цього завдання.

З огляду на викладене алгоритм ймовірнісного класифікатора у випадку двоелементної множини категорій може бути представлений блок-схемою (рис. 1).

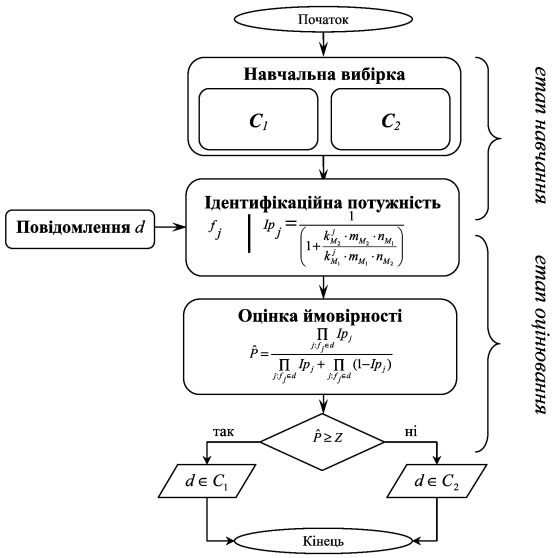

Рис. 1. Алгоритм ймовірнісного класифікатора у випадку двоелементної множини категорій

Для обґрунтування процедури вибору порогового значення наведемо результати експериментального дослідження. Так, навчальну вибірку утворюють текстові повідомлення, опубліковані в засобах масової інформації за період із червня 2006 року по березень 2009 року та розподілені шляхом експертного оцінювання (за участю 93 осіб) на дві категорії: 253 повідомлення категорії C_1 і 278 — категорії C_2. Зауважимо, що з 99% ймовірністю ця навчальна вибірка забезпечує репрезентативність з помилкою, що не перевищує 5,6 % [5, с. 313]:

$$\Delta=t\sqrt{\frac{w(1-w)}{n}\left(1-\frac{n}{N}\right)}=$$

$$=2{,}58\sqrt{\frac{0{,}5\cdot 0{,}5}{531}\left(1-\frac{531}{2^{32}}\right)}=0{,}056$$

Визначення помилки репрезентативності здійснювалось з урахуванням того, що про частку повідомлень однієї та іншої категорій в генеральній сукупності наперед нічого невідомо (w=0,5), а кількість повідомлень генеральної сукупності становить досить велике число ($N=2^{32}$).

Вибір порогового значення здійснюватимемо на основі порівняння ймовірностей

помилкового оцінювання автоматизованою системою повідомлень тестової вибірки. Позначимо помилку системи віднести повідомлення тестової вибірки до категорії C_1 в той час як воно насправді належить до C_2, як помилку 1-го типу, а помилку віднести повідомлення з категорії C_1 до категорії C_2 — як помилку 2-го типу. Відповідно загальна помилка оцінювання визначається як сума помилок 1-го та 2-го типу.

Тестову вибірку становили 70 текстів новин за період з червня 2006 року по березень 2009 року, розподілені експертами на дві категорії C_1 та C_2. Зафіксувавши згідно з рекомендаціями [4] як порогове значення $\hat{P} = 0{,}9$, загальна помилка оцінювання автоматизованою системою повідомлень тестової вибірки була 22,86%.

У процесі подальшого дослідження з'ясувалось, що причиною такої досить високої загальної помилки оцінювання повідомлень тестової вибірки є низька розподільча здатність навчальної вибірки. Тому подальші зусилля були зосереджені на підвищенні якості навчальної вибірки. Цей процес полягав у більш глибокому вивченні експертами структури повідомлень, які отримали помилкові оцінки, та встановленні більш строгих формальних критеріїв для розподілу повідомлень за категоріями.

У результаті отримано навчальну вибірку із 475 повідомлень (з рівнем довіри 99% помилка репрезентативності становить 5,9%). Відповідні показники помилки для неї та тестової вибірки при тому ж пороговому значенні наведено у табл. 1.

Таблиця 1

Показники помилок класифікації повідомлень тестової вибірки за різними навчальними вибірками

	Навчальна вибірка 531 пов.	**Навчальна вибірка 475 пов.**
Помилка 1 типу, %	10,00	15,71
Помилка 2 типу, %	12,86	2,86
Загальна помилка	22,86	18,57

Про покращення розподільчих властивостей 475-елементної навчальної вибірки свідчать, зокрема, box-and-whisker діаграми для обох навчальних вибірок та різниця між нижнім квартилем вибірки повідомлень категорії C_1 та верхнім квартилем вибірки повідомлень категорії C_2 (табл. 2, рис. 2 та 3).

Таблиця 2

Статистичні характеристики досліджуваних вибірок

Група повідомлень	**Показник**	**Навчальна вибірка 531 пов.**	**Навчальна вибірка 475 пов.**
Група повідомлень категорії C_1	Нижній квартиль, $\overset{\Lambda C_1}{P}_{0,75}$	0,811943	0,999991
	Медіана, $\overset{\Lambda C_1}{P}_{0,5}$	0,999552	1
	Верхній квартиль, $\overset{\Lambda C_1}{P}_{0,25}$	1	1
Група повідомлень категорії C_2	Нижній квартиль, $\overset{\Lambda C_2}{P}_{0,75}$	0	0
	Медіана, $\overset{\Lambda C_2}{P}_{0,5}$	0	0
	Верхній квартиль, $\overset{\Lambda C_2}{P}_{0,25}$	0,773597	0,697441

Однак із табл. 1 видно, що загальна помилка оцінювання залишилась все ще досить високою (18,57 %). Тому подальше експериментальне дослідження було зосереджено на зменшенні помилки оцінювання ймовірності за рахунок вибору найкращого порогового значення.

Ураховуючи властивості квартилей та особливості розподілу досліджуваної навчальної вибірки, логічно було б припустити, що порогове значення, при якому загальна помилка буде найменшою (а співвідношення помилок оцінювання 1 і 2 типу буде оптимальним), завжди знаходитиметься в інтервалі між верхнім квартилем групи по-

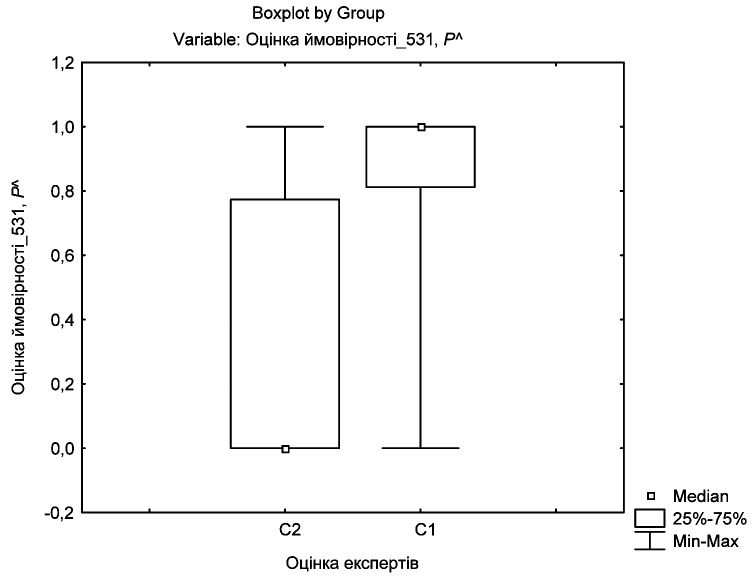

Рис. 2. Box-and-whisker діаграма для навчальної вибірки з 531 повідомленням

відомлень C_2 та нижнім квартилем групи повідомлень C_1. Причому, обираючи як порогове значення верхній квартиль повідомлень з C_2, мінімізується помилка 1-го типу. Обравши ж як порогове значення нижній квартиль повідомлень з C_1, гарантовано (з мінімальним рівнем помилки) відбиратимемо повідомлення, що належать до категорії C_1.

Підтвердженням таких припущень є рис. 4 та 5.

Вони відображають рівень помилки оцінювання залежно від порогового значення, отриманий у навчальній вибірці з 531 та 475 повідомленнями відповідно. Отже, оптимальне для обох типів помилок порогове значення знаходиться в інтервалі $\left[\overset{\Lambda\, C_2}{P}_{0,25};\ \overset{\Lambda\, C_1}{P}_{0,75}\right]$, при порогових значеннях, менших $\overset{\Lambda\, C_2}{P}_{0,25}$, зростає помилка 1 типу і зменшується помилка 2 типу, водночас при порогових значеннях, більших $\overset{\Lambda\, C_1}{P}_{0,75}$, зростає помилка 2 типу і зменшується помилка 1 типу.

З табл. 3 видно, що при пороговому значенні, рівному верхньому квартилю $\overset{\Lambda\, C_2}{P}_{0,25}$, мінімальною є помилка 2 типу, а при пороговому значенні, що дорівнює нижньому квартилю $\overset{\Lambda\, C_1}{P}_{0,75}$, помилка 1 типу.

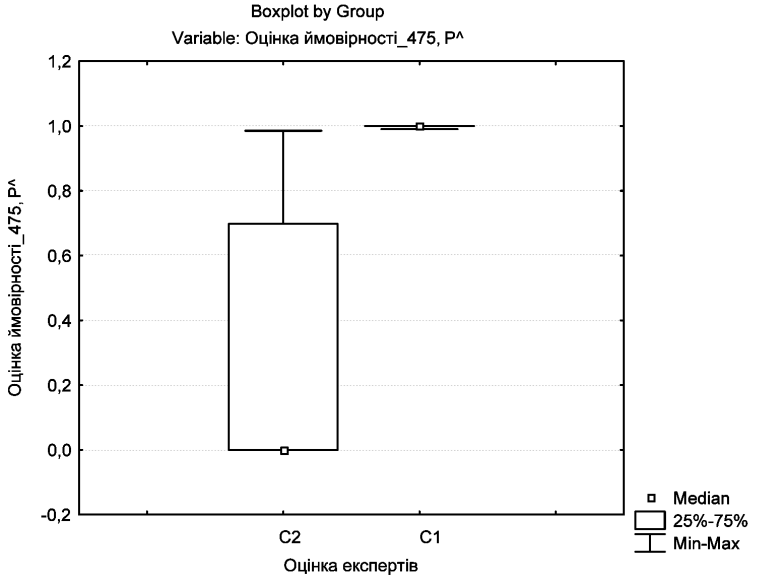

Рис. 3. Box-and-whisker діаграма для навчальної вибірки з 475 повідомленнями

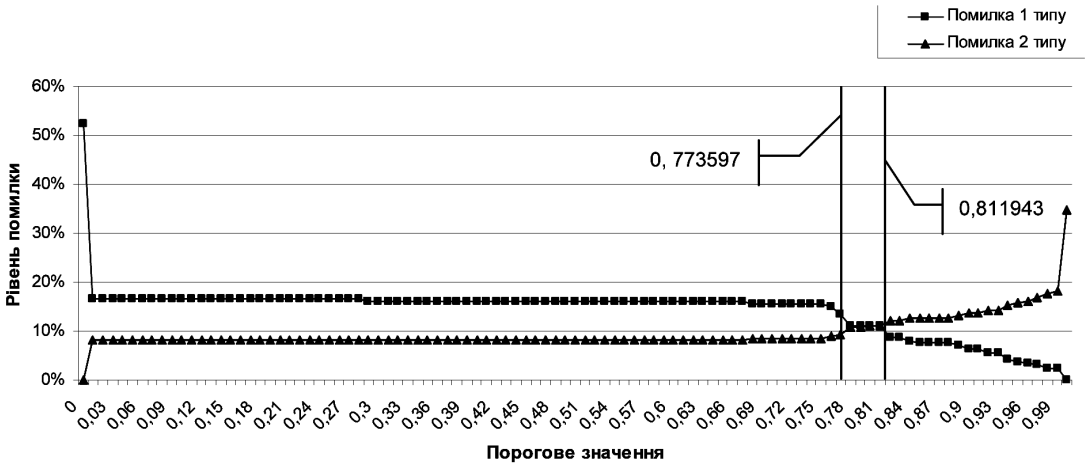

Рис. 4. Залежність помилки оцінювання від вибору порогового значення (у навчальній вибірці з 531 повідомленням).

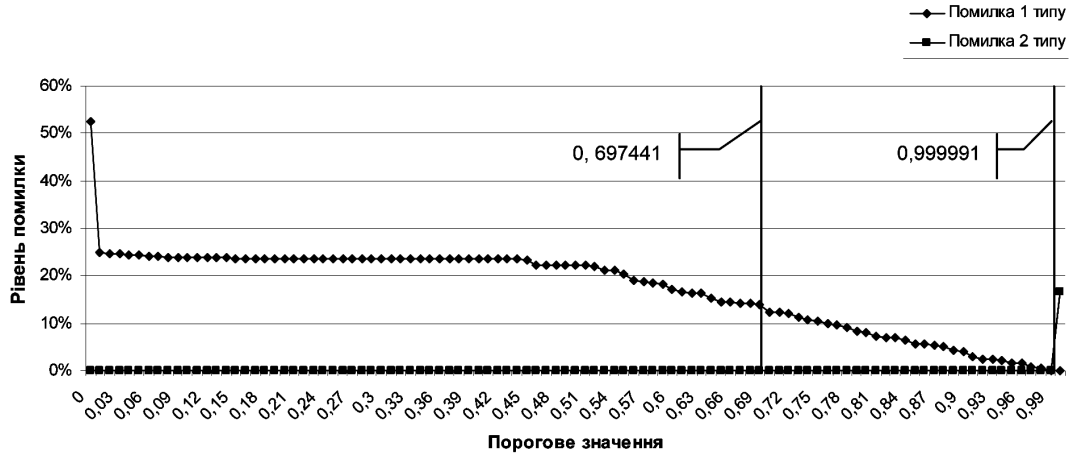

Рис. 5. Залежність помилки оцінювання від вибору порогового значення (у навчальній вибірці з 475 повідомленнями).

Таблиця 3

Показники помилки ймовірнісного класифікатора при різних порогових значеннях

	Навчальна вибірка (475 пов.)			**Тестова вибірка (70 пов.)**		
Порогове значення	$\overset{\Lambda}{P}{}^{C_2}_{0,25} = 0{,}697441$	$0{,}84 \in \left[\overset{\Lambda}{P}{}^{C_2}_{0,25};\ \overset{\Lambda}{P}{}^{C_1}_{0,75}\right]$	$\overset{\Lambda}{P}{}^{C_1}_{0,75} = 0{,}999991$	$\overset{\Lambda}{P}{}^{C_2}_{0,25} = 0{,}697441$	$0{,}84 \in \left[\overset{\Lambda}{P}{}^{C_2}_{0,25};\ \overset{\Lambda}{P}{}^{C_1}_{0,75}\right]$	$\overset{\Lambda}{P}{}^{C_1}_{0,75} = 0{,}999991$
Помилка 1 типу, %	12,28	5,95	0,00	27,14	20,00	2,86
Помилка 2 типу, %	0,00	0,00	10,75	1,43	1,43	14,29
Загальна помилка	12,28	5,95	10,75	28,57	21,43	17,14

На думку автора, можливість варіювання пороговим значенням є перевагою цього класифікатора, оскільки це дає змогу керувати точністю розподілу досліджуваних повідомлень: чим вище порогове значення, тим менше ймовірність помилково класифікувати повідомлення категорії C_2; чим нижче порогове значення, тим нижче ймовірність помилитися з класифікацією повідомлень категорії C_1. Встановивши формалізоване правило для вибору порогового значення, яке ґрунтується на статистичних даних, а не на досвіді та інтуїції дослідни-

ка, можемо забезпечити доповнення навчальної вибірки без залучення експертів, не погіршуючи розподільчих властивостей останньої. В цьому випадку можемо говорити про алгоритм ймовірнісного класифікатора як про систему із самонавчанням, адже він сам буде перебудовувати дані в процесі свого функціонування, і, таким чином, удосконалювати власну базу знань.

Викладене дає підстави дійти таких ***висновків:***

1. Алгоритм ймовірнісного класифікатора є чутливим до якості навчальної вибірки: чим виразніше відрізняються групи документів навчальної вибірки за обраною ознакою, тим вища точність алгоритму класифікації.
2. Для отримання точної відповіді ймовірнісного класифікатора у випадку двоелементної множини категорій шляхом встановлення порогових значень переходять від шкали [0,1] до шкали {0,1}.
3. Як порогове значення ймовірнісного класифікатора у випадку двоелементної множини категорій пропонується використовувати квартилі оцінки ймовірності, обчислювані за даними навчальної вибірки для кожної з двох груп повідомлень.
4. Визначена таким чином формалізована процедура вибору порогового значення дає підстави для подальших досліджень ймовірнісного класифікатора як системи із самонавчанням.

Література

1. Лифшиц Ю. Классификация текстов [Електронний ресурс] / Ю. Лифшиц // Современные задачи теоретической информатики: курс лекций. — 2005. — Режим доступу: http://yury.name/modern. **2. Sebastiani F.** Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys. — 2002. — № 1. — Vol. 34. — P. 1—47. **3. Ландэ Д. В.** Основы интеграции информационных потоков : моногр. / Д. В. Ланде. — К. : Инжиниринг, 2006. — 240 с. **4. Graham P.** Better Bayesian Filtering [Електронний ресурс] / Paul Graham // Paul Graham Site. — 2003. — Режим доступу: paulgraham.com/better.html. **5. Кремер Н. Ш.** Теория вероятностей и математическая статистика : учеб. / Н. Ш. Кремер. — М. : Юнити, 2002. — 543 с.

В статье приведено обоснование формализированной процедуры выбора порогового значения в задаче автоматизированной классификации текстов по двум категориям на базе вероятностной модели.

Ключевые слова: автоматизированная классификация текстов, наивный байесовский подход, пороговое значение, самообучающаяся система.

In the article the formalized procedure of threshold value choice in probabilistic model of the automated texts classification on two categories is proved.

Key words: automated text classification, naive Bayesian approach, threshold value, self-learning system.