Life and Other Mathematical Amusements. W.H. Freeman and Company, 1983. 4. Бандирська О.В. Стандартизація безнадлишкових рядів методом оптимальних структурних пропорцій: Автореф. канд. техн. Наук. – Львів: ДУЛП, 2000. 5. Різник В.В. Синтез оптимальних комбінаторних систем. – Львів: Вища шк., 1989. – 168 с.

УДК 519.765:519.767:004.93

Б. Павлишенко
Львівський національний університет імені Івана Франка

КЛАСИФІКАЦІЯ ПОВІДОМЛЕНЬ ГРУП НОВИН У ВЕКТОРНОМУ ПРОСТОРІ СЕМАНТИЧНИХ ПОЛІВ

Розглянуто класифікацію повідомлень груп новин у просторі семантичних полів. Проаналізовано ефективність баєсівського класифікатора та класифікатора за найближчими сусідами для різних навчальних та тестових вибірок повідомлень. Показано існування підмножини груп новин, для яких використання аналізованих класифікаторів є ефективним.

Ключові слова: інтелектуальний аналіз даних, класифікація текстів, векторна модель текстів, семантичні поля.

The classification of newsgroup messages in the space of semantic fields has been considered in this work. The effectiveness of Bayesian and nearest neighbors classifier for different training and test samples of messages has been analysed. The existence of a subset of newsgroups for which the use of analyzed classifiers is effective has been shown.

Key words: data mining, text classification, vector space model of texts, semantic fields.

Вступ

У роботах [1–3] наведені результати аналізу текстових масивів на основі концепції семантичних полів. Семантичні поля розглядають як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких є ефективним у задачах кластеризації та класифікації текстових документів. Однією із поширених моделей в інтелектуальному аналізі текстових даних є векторна модель, в якій текстові документи представляють у вигляді векторів у деякому фазовому просторі [4]. Базис цього простору утворюють частотні характеристики лексем. У роботі [1] розглянута теоретико-множинна концепція семантичних полів в масивах текстових даних. У роботі [2] запропонована модель кластеризації текстових документів у семантичному просторі, яка дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі істотно меншої розмірності, ніж у просторі, утвореному лексемним складом текстової вибірки. У задачах аналізу текстового змісту актуальними є теорії лексичної семантики, зокрема, вчення про семантичні поля. Спорідненими об'єктами у комп'ютерній інформатиці є семантичні мережі, в яких відображаються змістовні зв'язки між різними концептами. Одним із прикладів ієрархічно-організованої семантичної мережі можна розглядати систему WordNet, яка розроблена у Прінстонському університеті [5]. Лексемний склад в цій системі організований у вигляді синсетів, під якими розуміють набори лексем синонімічного ряду, які є взаємозамінними у заданих контекстах. Бази даних WordNet створили експерти-лексикографи. Іменники, дієслова, прикметники та прислівники організовані у синсети – множини синонімів. Іменники та дієслова

згруповані відповідно до семантичних полів. Семантична структурна організація лексемного складу словника може бути використана у відповідних алгоритмах класифікації та кластеризації текстових об'єктів з точки зору зменшення розмірності задач аналізу та виявлення нових семантичних зв'язків в онтології предметної області, до якої відносять аналізований масив текстів. У роботі [6] введено поняття семантичного домена, який описує деяку семантичну область розгляду тої чи іншої теми обговорень, наприклад, економіка, політика, фізика, програмування тощо. Для розгляду алгоритмів текстової кластеризації часто використовують стандартизовані масиви текстових документів. Однією із таких колекцій є 20-Newsgroups [http://qwone.com/~jason/20Newsgroups/], яка включає у себе колекцію приблизно 20 тисяч документів близько 20 груп новин. Цю колекцію використовують у тестових задачах інтелектуального аналізу текстових масивів, зокрема у задачах класифікації та категоризації текстових масивів.

Постановка задачі

Розглянемо класифікацію текстових документів у просторі семантичних полів. Для порівняння розглянемо байесівський класифікатор та класифікатор за найближчими k сусідами. Як навчальну та тестову вибірки використаємо повідомлення груп новин стандартизованої текстової бази даних 20NewsGroups. Проаналізуємо кількісні характеристики ефективності класифікаторів із різними параметрами текстових вибірок.

Класифікатори текстових документів у просторі семантичних полів. Сукупність текстових документів опишемо такою множиною:

$$D = \{d_j \mid j = 0,1,2...,N_d\}. \quad (1)$$

Введемо множину семантичних полів

$$S = \{s_k \mid k = 1,2...,N_s\}. \quad (2)$$

Під семантичним полем розуміють таку множину лексем, у якій вони об'єднані певним спільним поняттям [1, 5, 6]. Прикладом семантичних полів може бути поле руху, поле комунікації, поле сприйняття та інші. Нехай існує певний словник лексем, які зустрічаються у текстових масивах $W = \{w_i \mid i = 1,2...,N_w\}$. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \overset{U_{ws}}{\rightarrow} s_k, i = 1,2...,N_w \right\}. \quad (3)$$

Введемо частоту семантичного поля p_{kj}^{sd} за такою формулою:

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \quad f_s(w_i, s_k) = \begin{cases} 1, w_i \in W_k^s \\ 0, w_i \notin W_k^s \end{cases}, \quad (4)$$

де p_{ij}^{wd} т – текстова частота лексеми w_i в документі d_j, яка визначається відношенням наявної кількості лексеми w_i до загальної кількості лексем у документі d_j. Сукупність значень p_{kj}^{sd} утворюють матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,j=1}^{N_s,N_d}. \quad (5)$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right) \quad (6)$$

відображає документ d_j в N_s-мірному просторі текстових документів із базисом, утвореним семантичними полями.

Нехай існують деякі категорії текстових документів. Ці категорії можуть мати різну природу, наприклад, можуть визначати авторський ідеолект, дискурс, характеризувати різні об'єкти, явища, події тощо. У нашому аналізі такі категорії утворюють групи новин. Множину цих категорій позначимо

$$Categories = \{ Ctg_m \mid m = 1,2,...,N_{ctg} \}, \quad (7)$$

де $N_{ctg} = |Categories|$ визначає розмір множини категорій. За даними категоріями розподілені текстові документи множини D (1). Завдання полягає у пошуку цільової функції, яка описується відображенням

$$F_{d \to ctg} : Categories \times D \to \{0,1\} \tag{8}$$

Розглянемо наївний байесівський класифікатор текстових локументів. У відомих методах текстової класифікації на основі наївної байесівського класифікатора використовують представлення документів за допомогою частот відповідних ключових слів [7, 8]. Підхід, який базується на представленні документів частотними характеристиками семантичних полів, є перспективним внаслідок меншої розмірності семантичного фазового простору. Знайдемо апостеріорну ймовірність того, що за деяким набором частот семантичних полів документ d_j належить до категорії ctg_m. За теоремою Байеса визначимо

$$P(ctg_m \mid p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}) = \frac{P(ctg_m) P(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \mid ctg_m)}{P(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd})} \tag{9}$$

У реалізації наївного байесівського класифікатора роблять істотне припущення про умовну незалежність ознак об'єктів [7,8]. У такому випадку умовну ймовірність $P(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \mid ctg_m)$ апроксимують добутком умовних ймовірностей $P(p_{ij}^{sd} \mid ctg_m)$. Неперервні розподіли $P(p_{ij}^{sd} \mid ctg_m)$ часто апроксимують нормальним гауссовим розподілом. Параметрами цього розподілу розглядають математичне сподівання та дисперсію семантичних полів. Доповненням до розрахунку наївного байесівського класифікатора є правило прийняття рішень про віднесення аналізованого документа до тієї чи іншої категорії [7,8]. У найпростішому випадку таке правило може приймати рішення про належність документа до заданої категорії, якщо розрахована апостеріорна ймовірність для такої категорії при заданих частотах семантичних полів є найбільшою, тобто

$$\begin{aligned} &Category(d_j) = ctg_m : P(ctg_m \mid p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}) = \\ &= max\left\{P(ctg_k \mid p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}) \mid k = 1,2,...,N_{ctg}\right\} \end{aligned} \tag{10}$$

Ймовірності $P(p_{ij}^{sd} \mid ctg_m)$ формують на деякому навчальному категоризованому масиві текстових документів.

Розглянемо класифікацію за найближчими k сусідами, яку називають kNN класифікацією [7, 8, 9]. Цей метод зараховують до векторних класифікаторів. В основі векторних методів класифікації лежить гіпотеза компактності. Згідно із цією гіпотезою, документи, які належать одному і тому ж класу утворюють компактну область, а області, які належать різним класам не перетинаються. Як міру близькості між документами виберемо евклідову відстань. Поряд із евклідовою відстанню часто використовують значення косинуса кута між двома векторами. У kNN класифікації межі категорій визначають локально. Деякий документ зараховують до категорії, яка є домінуючою для k його сусідів. У випадку $k=1$ документу приписують категорію його найближчого сусіда. Згідно із гіпотезою компактності тестовий документ d має ту категорію, яку мають більшість документів навчальної вибірки у деякому просторовому локальному околі документа d.

Розглянемо оцінки точності класифікації документів. Прийняття рішення класифікатором про належність документа d_i до категорії ctg_j позначимо $Class(d_i) = Ctg_j$. Множина документів, які визначені класифікатором як належні до категорії ctg_j і які дійсно належать цій категорії згідно з експертною оцінкою має вигляд

$$Set_1^{tclass} = \{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\}. \tag{11}$$

Множина документів, які визначені класифікатором як належні до категорії ctg_j, має вигляд

$$Set_2^{tclass} = \{d_i \mid Class(d_i) = Ctg_j\}. \tag{12}$$

Множина документів, які належать до категорії ctg_j має вигляд

$$Set_3^{tclass} = \{d_i \mid d_i \in Ctg_j\}. \quad (13)$$

Кількості елементів у множинах $Set_1^{tclass}, Set_2^{tclass}, Set_3^{tclass}$ визначаються кардинальними числами цих множин $|Set_1^{tclass}|, |Set_2^{tclass}|, |Set_3^{tclass}|$. Для характеристики класифікаторів використовують поняття точності (precision) та повноти (recall) [8,9]. Точність класифікатора визначають як відношення кількості елементів множини Set_1^{tclass} до елементів множини Set_2^{tclass}

$$Pr_j^{tclass} = \frac{Set_1^{tclass}}{Set_2^{tclass}} = \frac{\left|\{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\}\right|}{\left|\{d_i \mid Class(d_i) = Ctg_j\}\right|} \quad (14)$$

Повноту визначають як відношення кількості елементів множини Set_1^{tclass} до елементів множини Set_3^{tclass}

$$Rc_j^{tclass} = \frac{Set_1^{tclass}}{Set_3^{tclass}} = \frac{\left|\{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\}\right|}{\left|\{d_i \mid d_i \in Ctg_j\}\right|} \quad (15)$$

Індекс j у характеристиках Pr_j^{tclass}, Rc_j^{tclass} визначає категорію, а індекс *tclass* визначає тип класифікатора. У наших дослідженнях

$$tclass = \{NB, nKNN\} \quad (16)$$

Кожна категорія документів характеризується своїми значеннями Pr_j^{tclass}, Rc_j^{tclass}. Для загальної характеристики класифікатора знайдемо макроусереднення показників Pr_j^{tclass}, Rc_j^{tclass} за усіма категоріями

$$Pr_{mean}^{tclass} = \frac{1}{N_{ctg}} \sum_{i=1}^{N_{ctg}} Pr_i^{tclass}, \quad (17)$$

$$Rc_{mean}^{tclass} = \frac{1}{N_{ctg}} \sum_{i=1}^{N_{ctg}} Rc_i^{tclass}. \quad (18)$$

Розкид показників Pr_j^{tclass}, Rc_j^{tclass} за категоріями охарактеризуємо середньоквадратичним відхиленням

$$Pr_{std}^{tclass} = \sqrt{\frac{1}{N_{ctg}} \sum_{i=1}^{N_{ctg}} (Pr_i^{tclass} - Pr_{mean}^{tclass})^2}, \quad (19)$$

$$Rc_{std}^{tclass} = \sqrt{\frac{1}{N_{ctg}} \sum_{i=1}^{N_{ctg}} (Rc_i^{tclass} - Rc_{mean}^{tclass})^2}. \quad (20)$$

Очевидно, що чим більша кількість можливих категорій кластеризації, тим складніше завдання є перед класифікатором документів і тим більша ймовірність похибки, тому необхідно ввести деяку кількісну характеристику, яка б показувала ефективність класифікатора, враховуючи кількість категорій. Як одну із таких характеристик розглянемо *покращення (improvement),* яку визначемо як відношення повноти до частоти випадкового визначення правильної категорії

$$Impr^{tclass} = \frac{Rc_{mean}^{tclass}}{P_{ctg}^{prb}}. \quad (21)$$

Покращення $Impr^{tclass}$ характеризує ефективність класифікатора порівняно із випадковим вибором категорії. Для простоти розрахунків вважаємо, що ймовірність правильного випадкового вибору категорії документа є однакова для всіх категорій

$$P_{ctg}^{prb} = \frac{1}{|Categories|}. \quad (22)$$

Тоді отримаємо

$$Impr^{tclass} = |Categories| \cdot Rc_{mean}^{tclass}. \quad (23)$$

Експериментальна частина

Для експериментального вивчення класифікації текстових документів у просторі семантичних полів ми вибрали стандартизовану текстову базу повідомлень 20NewsGroups [http://qwone.com/~jason/20Newsgroups/]. Ця база містить близько 20000 повідомлень, які рівномірно розподілені по 20 групах новин. Для формування семантичного простору вибрано лексеми, згруповані за семантичними полями іменників та дієслів у семантичній мережі WordNet [5]. Семантичні полі у мережі WordNet (http://wordnet.princeton.edu) представлені лексикографічними файлами. У наших дослідженнях ми використали семантичні поля іменників та дієслів. Семантичні поля іменників складаються із 26 лексикографічних файлів, із яких ми відібрали 54464 лексеми. Семантичні поля дієслів містять 15 лексикографічних файлів, у які ми відібрали 9097 лексем. У семантичні поля також ввійшли похідні форми лексем. За допомогою розробленого програмного забезпечення здійснена початкова обробка текстового масиву, вилучено допоміжні символи та текстові елементи, які не несуть семантичної інформації. Для кожного документа та вибірки в цілому, обраховано частотні словники, на основі яких розраховано матрицю M_{sd} (5) типу документ-частота_семантичного _поля. На основі цієї матриці ми досліджували два типи класифікатрорів – наївний байесівський класифікатор та класифікатор за найближчими сусідами. У дослідженнях розглядались різні параметри класифікації, зокрема, досліджувалась класифікація, у якій навчальна та тестова вибірки документів збігалися та випадки, коли вони були різні. Також досліджувався випадок об'єднаної множини семантичних полів та випадки класифікації лише за полями іменників чи полями дієслів. Досліджувались випадки класифікації текстової вибірки із вилученими високочастотними лексемами, які разом становлять 50% текстового наповнення масиву документів. Також проаналізовані випадки класифікації документів окремих категорій, які в результаті класифікаційного аналізу виявились найбільш категоріально виразними у просторі семантичних полів. Розглянемо основні отримані результати. Точність та повнота для категорій у випадку байесівської класифікації при збігу навчальної та тестової вибірки об'ємом 20000 повідомлень множин для 40 семантичних полів та 20 категорій наведена на рис. 1. Такі ж розрахунки для класифікатора за найближчими k сусідами при $k=3$ наведено на рис. 2. На цих рисунках відображені категоріальні розподіли точності та повноти. У випадку байесівського класифікатора для деяких категорій показники точності та повноти є задовільними, а для деяких є низькими. Метод класифікації за найближчими сусідами показує високу ефективність. При класифікації тестового документа, який одночасно належить навчальній вибірці відбувається попадання цього документі у його індивідуальний семантичний окіл. При розгляді семантичного підпростору, утвореного лише семантичними полями іменників чи дієслів, точність розрахунків зменшується. При збільшенні кількості сусідів втрачається точність класифікатора. Це свідчить про високу щільність текстових документів різних категорій у семантичному просторі і впливає на точність визначення категорії документа. При $k=1$ точність та повнота зростають. Це означає, що кожний документ у семантичному просторі заходиться у окремому місці і не відбувається перекриття семантичних околів для різних документів.

. Результати макроусереднення показників класифікації по групах новин при різних умовах наведені у таблиці. Низькі значення точності та повноти байесівської класифікації повідомлень для деяких груп новин можуть бути зумовлені тим, що хоч повідомлення і належать декій групі, однак можуть містити нейтральну текстову інформацію щодо основної тематики групи.

Розглянемо випадок розділення текстової вибірки на навчальну та тестову так, щоб документи із цих двох вибірок не збігались. Документи у вибірці розміщувались у випадковому порядку. Слід відмітити, що для одних і тих самих навчальних та тестових вибірок та категорій спостерігаються різні значення точності та повноти при використанні байесівського класифікатора та класифікатора за найближчими сусідами. Метод найближчих сусідів дає високі результати

точності та повноти класифікації у випадку коли тестова вибірка є підмножиною навчальної вибірки. Коли ці вибірки є відмінні, точність класифікатора за найближчими сусідами є суттєво меншою, однак є більшою за точність байесівського класифікатора у цих самих умовах. Невисокі значення точності та повноти при відмінності навчальної та тестової вибірки можуть пояснюватись специфікою текстової бази груп новин.

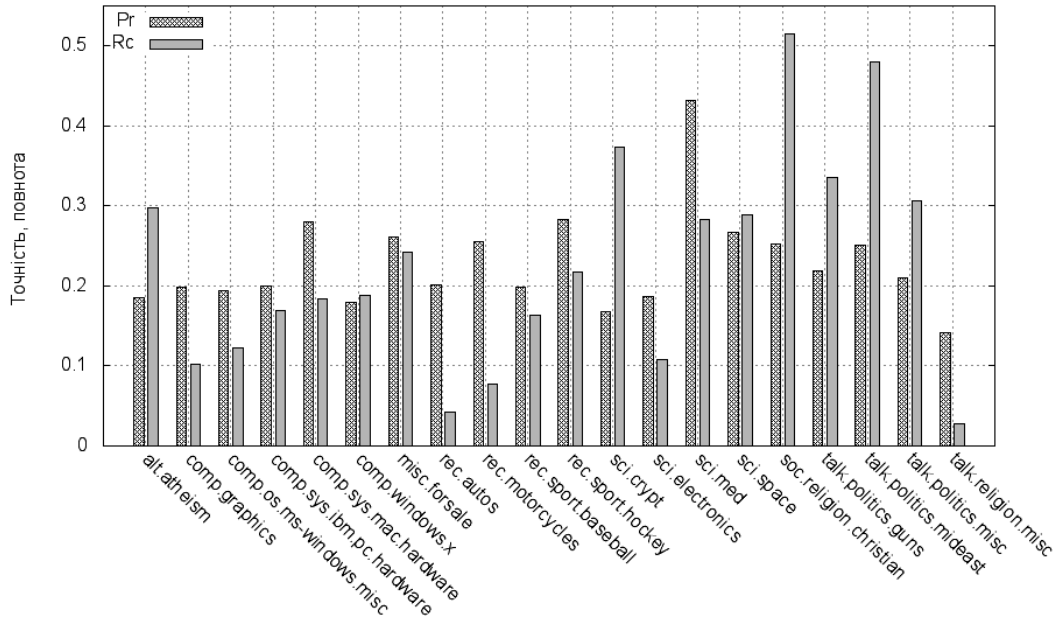

Рис. 1. Точність та повнота байесівського класифікатора для груп новин

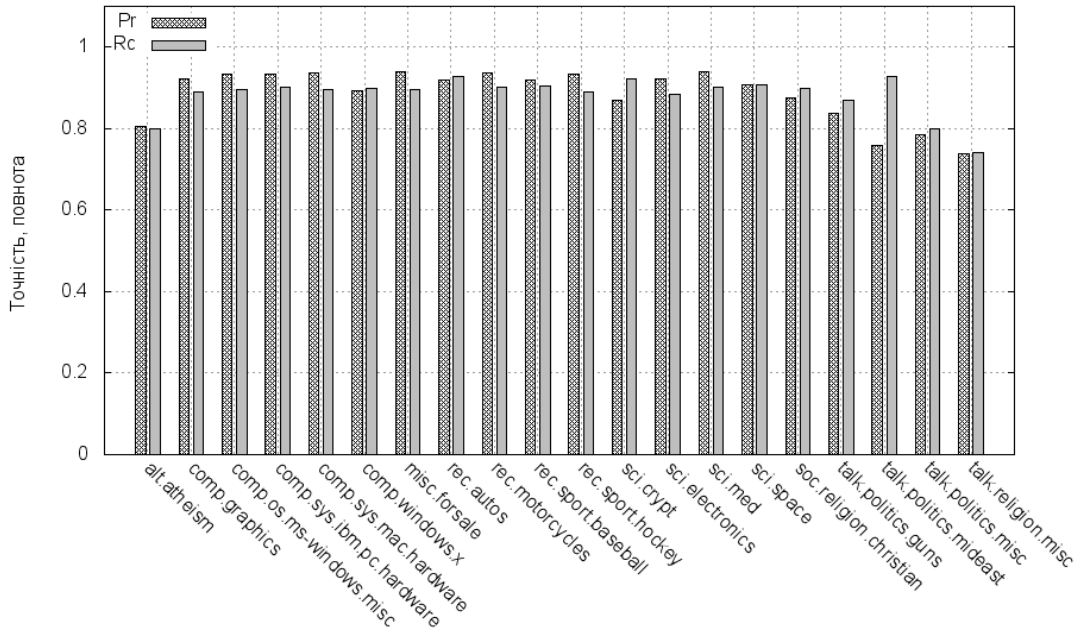

Рис. 2. Точність та повнота класифікатора за найближчими k сусідами (k=3) для груп новин

Насамперед це малий об’єм текстової інформації у кожному повідомленні, яка у середньому становить близько 1Кбайт. Це впливає на якість представлення текстових повідомлень у просторі семантичних полів. Наявність повідомлення у групі новин відображає членство у цій групі автора повідомлення і не завжди є релевантим до тематики, тому може не належати області у просторі семантичних полів, яка відображає тематику групи. Для випадку, коли навчальна та тестова вибірки

збігаються, тоді кожний документ є представлений у деякому семантичному околі простору і при $k=1$ і цей документ попадає у свою ж область, що пояснює високу точність класифікатора kNN у цих умовах. Очевидно, що широковживані лексеми несуть мінімальну семантичну інформацію і часто виконують допоміжну роль у тексті. Розглянемо частотний словник лексем, упорядкований за частотою лексем, які сумарно становлять 50 % лексемного наповнення текстів. Для аналізованої вибірки такий список містить 100 лексем. На основі цього списку утворимо лексемний фільтр, за допомогою якого виділимо ці лексеми із текстової вибірки. Для відфільтрованого текстового масиву побудуємо матрицю семантичних векторів документів і застосуємо класифікатор за найближчими сусідами. Отримано усереднені значення точності та повноти для розділених навчальних та тестових вибірок ($\left|D_{set}^{training}\right|$=15000, $\left|D_{set}^{test}\right| = 1/3\left|D_{set}^{training}\right|$):

$$Pr_{mean}^{tclass} = 0.31\,, Rc_{mean}^{tclass} = 0.30\;\; Impr^{tclass} = 6.12\,.$$

Показники класифікаторів при різних параметрах класифікації

Класифікатори	Pr_{mean}^{tclass}	Pr_{std}^{tclass}	Rc_{mean}^{tclass}	Rc_{std}^{tclass}	$Impr^{tclass}$
Умови класифікації: $D_{set}^{test} = D_{set}^{training}$, $S = \{ s_k \mid (\forall w_i \in s_k : w_i \in Nouns) \vee (\forall w_i \in s_k : w_i \in Verbs) \}$					
NB	0.2283	0.2262	0.0621	0.1349	4.5237
kNN (k=1)	0.9731	0.0494	0.9729	0.0482	19.4589
kNN (k=3)	0.8855	0.0653	0.8829	0.0477	17.6586
kNN (k=10)	0.5500	0.0762	0.5335	0.0856	10.6696
Умови класифікації: $D_{set}^{test} = D_{set}^{training}$, $S = \{ s_k \mid (\forall w_i \in s_k : w_i \in Nouns) \}$					
NB	0.1863	0.1765	0.0701	0.1281	3.5295
kNN (k=1)	0.4957	0.0628	0.4870	0.0752	9.7395
Умови класифікації: $D_{set}^{test} = D_{set}^{training}$, $S = \{ s_k \mid (\forall w_i \in s_k : w_i \in Verbs) \}$					
NB	0.1673	0.1473	0.0608	0.1337	2.9464
kNN (k=1)	0.9713	0.0496	0.9710	0.0470	19.4209
Умови класифікації: $D_{set}^{test} \notin D_{set}^{training}$, $\left|D_{set}^{training}\right|$=15000, $\left|D_{set}^{test}\right| = 5000$ $S = \{ s_k \mid (\forall w_i \in s_k : w_i \in Nouns) \vee (\forall w_i \in s_k : w_i \in Verbs) \}$					
NB	0.2177	0.0718	0.2168	0.1374	4.3497
kNN (k=1)	0.2877	0.0589	0.2840	0.0791	5.6663
kNN (k=3)	0.2883	0.0606	0.2853	0.0875	5.6863
Умови класифікації: $D_{set}^{test} \notin D_{set}^{training}$, $\left|D_{set}^{training}\right| = 5000$ $\left|D_{set}^{test}\right| \approx 5000$, $S = \{ s_k \mid (\forall w_i \in s_k : w_i \in Nouns) \vee (\forall w_i \in s_k : w_i \in Verbs) \}$					
NB	0.1957	0.0592	0.2007	0.1269	4.0696
kNN (k=1)	0.2154	0.0484	0.2131	0.0752	4.2577
kNN (k=3)	0.2268	0.0566	0.2282	0.1012	4.5618

Як випливає із отриманих результатів вилучення широковживаних лексем, які становлять 50 % наповнення текстових масивів покращує точність та повноту класифікатора і істотно зменшує обсяг обчислень у класифікаційному аналізі внаслідок зменшення текстової вибірки наполовину лексем.

При формуванні класифікаційного простору на основі підібраних для даної вибірки ключових слів можна досягнути значно кращих результатів. Однак формування такої вибірки потребує

значних експертних ресурсів, крім того розмірність класифікаційного простору значно зросте. Така вибірка буде спеціалізовано лише для даного типу тематики текстової колекції. В той же час байесівська класифікація у просторі семантичних полів є універсальною для всіх типів текстових колекцій і не потребує додаткових налаштувань. Байеівський класифікатор характеризується великим значенням середньоквадратичного відхилення по класифікаційним категоріям. Це свідчить про те, що для деяких категорій він дає задовільні результати.

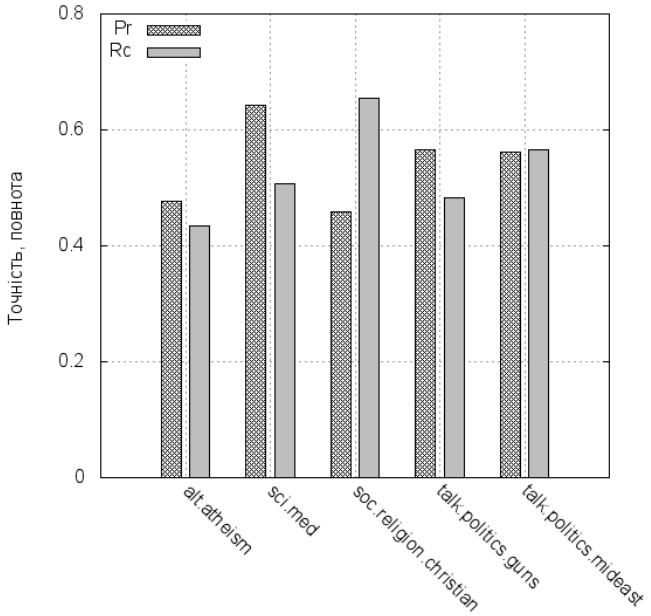

Рис. 3. Показники байесіської класифікації для вибраних семантично виразних категорій

Не зважаючи на низькі значення макроусередненої точності та повноти байесівського класифікатора при великій кількості категорій для аналізованого типу текстових документів цей метод може бути використаний у попередньому опрацюванні колекції текстових документів, зокрема тієї частки категорій, на яких цей класифікатор дає задовільні результати. Очевидно, що байесівський класифікатор може бути використаний для деякої підмножини груп новин, повідомлення яких є семантично виразними по відношенню до інших категорій. На рис. 3 зображена точність та повнота байесівського класифікатора для підмножини семантично виразних груп новин у випадку роздільної навчальної та тестової вибірки. Макроусереднені показники мають такі значення $Pr_{mean}^{tclass} = 0.5407$, $Rc_{mean}^{tclass} = 0.5290$. Отримані результати свідчать про те, що у множині груп новин існує підгрупа новин, повідомлення якої можуть бути ефективно класифіковані байесівським класифікатором.

Висновки

Проаналізовано можливість використання наївного баєсівського класифікатора (NB) та класифікатора за найближчими сусідами (kNN) у класифікаційному семантичному аналізі повідомлень груп новин. Текстові повідомлення розглянуті у векторному просторі, базис якого утворюють частотні характеристики семантичних полів іменників та дієслів. Розмірність такого базису є істотно меншою ніж у випадку широковживаної класифікації текстів за ключовими словами. Виявлено високу ефективність kNN класифікатора у випадку збігу навчальної та тестової вибірки. Виявлено підмножину групп новин на яких NB класифікатор дає задовільні результати. Категоріальні розподіли точності та повноти можуть істотно відрізнятись для різних класифікаторів при одних і тих самих навчальних та тестових вибірках. Отримані результати свідчать про ефективність реалізації NB та kNN класифікації у просторі семантичних полів і відображають сукупність характеристик розглянутих класифікаторів та текстової вибірки заданого типу повідомлень груп новин. Для іншого типу вибірок кількісні характеристики

семантичних класифікаторів можуть істотно відрізнятись. У наступних дослідженнях ми плануємо розглянути класифікацію текстів у семантичному просторі для вибірок авторських текстів у задачах аналізу авторського ідеолекта, а також інші стандартизовані текстові вибірки.

1. Павлишенко Б. М. Використання концепції семантичного поля у векторній моделі текстових документів // Східно-Європейський журнал передових технологій. – 2011. – № 6/2(54). – С. 7–11. 2. Павлишенко Б. М. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів // Електроніка та інформаційні технології. –2011. – Вип. 1. – С .212–222. 3. Павлишенко Б. М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. – 2012. – № 1. – С. 69–76. 4. Pantel Patrick, Turney Peter D. From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research.–2010.–vol.37.–pp.141-188. 5. Fellbaum C. WordNet. An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998, 432p. 6. Gliozzo Alfio, Strapparava Carlo. Semantic Domains in Computational Linguistics. Springer, 2009 – 132 p. 7. Брасегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов: учеб. Пособие.– СПб.: БХВ–Петербург,2009.-512с.:ил. 8. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, № 1. – pp. 1–47. 9. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. – 496 p.

УДК 004.421.2:517.443

І. Процько
Національний університет “Львівська політехніка”,
кафедра систем автоматизованого проектування

СИНТЕЗ ТА ОБЧИСЛЕННЯ ОСНОВНИХ ТИПІВ ДПХ НА ОСНОВІ ЦИКЛІЧНИХ ЗГОРТОК

Розглянуто підхід до ефективного обчислення основних чотирьох типів дискретного перетворення Хартлі (ДПХ) на основі циклічних згорток. Параметри твірного масиву базисної квадратної матриці використано для синтезу алгоритму.

Ключові слова: дискретні перетворення Хартлі, твірний масив, синтез алгоритму, циклічна згортка.

The general method of efficient computation four types discrete Hartley transform using of circular convolutions is considered. The parameters of hash array of basis square matrix for algorithm synthesis are used.

Key words: discrete Hartley transforms, hash array, algorithm synthesis, cyclic convolution.

Вступ

Для опису даних в їх спектральному гармонічному образі застосовуються високоефективні дискретні перетворення класу Фур’є. У більшості застосувань опрацьовують інформацію над послідовностями дійсних даних. Тому обробка ДПФ над дійсними даними є інформаційно надлишкова, а саме дійсна частина ДПФ є парною функцією та уявна непарною [1]. Одну з