УДК 519.765:519.767:004.93

Б.М. ПАВЛИШЕНКО*

ФОРМУВАННЯ БАЗИСУ СЕМАНТИЧНОГО ПРОСТОРУ ТЕКСТОВИХ ДОКУМЕНТІВ ЗА ДОПОМОГОЮ ГЕНЕТИЧНИХ АЛГОРИТМІВ

*Львівський національний університет імені Івана Франка, Львів, Україна

Анотація. *У роботі розглянуто використання генетичних алгоритмів для формування набору семантичних полів, частотні характеристики яких утворюють базис семантичного простору в інтелектуальному аналізі текстових даних. На прикладі класифікаційного аналізу текстових повідомлень груп новин показана ефективність генетичної оптимізації семантичного базису векторного простору текстових документів.*
Ключові слова: *генетичні алгоритми, інтелектуальний аналіз даних, класифікація текстів, семантичні поля.*

Аннотация. *В работе рассмотрено использование генетических алгоритмов для формирования набора семантических полей, частотные характеристики которых образуют базис семантического пространства в интеллектуальном анализе текстовых данных. На примере классификационного анализа текстовых сообщений групп новостей показана эффективность генетической оптимизации семантического базиса векторного пространства текстовых документов.*
Ключевые слова: *генетические алгоритмы, интеллектуальный анализ данных, классификация текстов, семантические поля.*

Abstract. *Genetic algorithms usage for forming set of semantic fields has been discussed in this paper. These fields form the semantic space basis by their frequency characteristics in data mining of text documents. The classification analysis of newsgroups texts messages shows the effectiveness of genetic optimization of semantic basis of vector space for text documents.*
Keywords: *genetic algorithms, data mining, texts classification, semantic fields.*

1. Вступ

В інтелектуальному аналізі текстових даних часто використовують векторну модель текстових документів. 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. Однією із актуальних задач є пошук оптимальних векторних підпросторів документів для класифікаційного та кластерного аналізу текстових документів. Зокрема, задача полягає у відборі семантичних полів, частотні характеристики яких можуть використовуватись як вхідні параметри текстових класифікаторів із задовільною точністю. Розв'язок такої задачі оптимізує необхідну кількість обрахунків та точність класифікатора в інтелектуальному аналізі текстів. Один із перспективних методів формування базису семантичного простору може бути побудований із використанням генетичних алгоритмів.

2. Аналіз останніх досліджень

У роботі [1] розглянута теоретико-множинна концепція семантичних полів у 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 суттєво меншої розмірності, ніж у

просторі, утвореному частотними характеристиками лексемного складу 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 відображено змістовні зв'язки між різними концептами. Одним із прикладів ієрархічної семантичної мережі можна розглядати систему WordNet, яка розроблена у Прінстонському університеті [5]. Лексемний склад у цій системі організований у вигляді синсетів, під якими розуміють набори лексем синонімічного ряду, що є взаємозамінними у заданих контекстах. Іменники та дієслова згруповані відповідно до семантичних полів. У роботі [6] запропонована концепція семантичних доменів, яка доповнює теорію семантичних полів. Визначення семантичних доменів базується на відповідних текстових колекціях, які належать до аналізованого домена і характеризують семантичні поняття, що виокремлюють аналізований домен. Семантична структурна огранізація лексемного складу словника може бути використана у відповідних алгоритмах класифікації та кластеризації текстових об'єктів для формування ефективного базису векторного простору текстових документів. 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.

3. Мета статті

Розглянемо генетичні алгоритми оптимізації з точки зору формування ефективного базису на основі семантичних полів для векторного простору текстових документів. Створимо векторну модель генетичного відбору семантичних полів у задачі класифікації текстових документів. Експериментально дослідимо формування семантичного підпростору у класифікаційному аналізі на прикладі класифікатора за найближчими k сусідами. Як основу для експериментального дослідження виберемо стандартизовану колекцію повідомлень груп новин 20-Newsgroups.

4. Основні положення теорії генетичної оптимізації параметрів

Розглянемо можливість використання еволюційного програмування та генетичних алгоритмів для формування оптимального простору семантичних полів у задачах інтелектуального аналізу текстів. Опис генетичних алгоритмів можна знайти у роботах [7–12]. Генетичні алгоритми використовують у широкому класі оптимізаційних задач, які полягають у пошуку набору вхідних параметрів, мінімізуючих деяку цільову функцію. Як цільову функцію можна розглядати похибку класифікатора або деяку кількісну характеристику кластерної структури текстових документів. Як вхідні параметри оптимізаційної задачі розглянемо набір семантичних полів, що утворюють базис векторного простору текстових документів. Ідея генетичних алгоритмів була запропонована Джоном Холандом у 1975 році у Мічіганському університеті і полягає у використанні основних положень еволюційної теорії Дарвіна, зокрема, принципу природного відбору та спадкової мінливості у розв'язку оптимізаційних задач. Розглянемо основні положення генетичних алгоритмів у контексті задачі пошуку оптимального семантичного базису для інтелектуального аналізу текстових документів, зокрема, на основі класифікаційних алгоритмів. Генетичний алгоритм пошуку оптимального семантичного базису розглянемо за класичною схемою. Набір із вхідних параметрів називають хромосомою або особою. У простому випадку особа утворена на основі однієї хромосоми. Сукупність хромосом утворює популяцію. Сукупність семантичних полів векторного базису у контексті генетичних алгоритмів назвемо хромосомою семантичних полів. У генетичних алгоритмах базовими є оператори

вибору батьківських хромосом для утворення нових хромосом у наступних поколіннях, оператор кросовера для рекомбінації хромосом, оператор мутації, оператор селекції хромосом для їх відбору у наступне покоління. Один із поширених і простих операторів відбору батьків для породження нової хромосоми називається панмекією і полягає у випадковому виборі батьківської пари. Характерною особливістю такого підходу є розмір популяції. Такий метод ефективний при генерації популяцій великих розмірів. Оператор відбору, який називається інбридингом, полягає у відборі близьких за відстанню Хемінга батьківських хромосом, а оператор аутбридинг полягає у відборі найвіддаленіших осіб. Часто використовують оператор випадкового однорідного відбору хромосом, який має таку схему. Хромосоми розміщують на одній лінії, причому кожній хромосомі відводять відрізок лінії, пропорційний їхній пристосованості у популяції або, наприклад, обернено пропорційний цільовій функції, якщо задача полягає у мінімізації цільової функції. Далі на основі рівномірного розподілу вибирають точки на цій лінії. Батьківською хромосомою стає та хромосома, на відрізку якої знаходиться вибрана точка.

Очевидно, що при такому підході хромосоми із найменшими значеннями цільової функції найчастіше стають батьківськими хромосомами, які породжують дочірні хромосоми наступного покоління популяції. Оператор селекції виконує відбір хромосом у наступне покоління на основі кількісних значень цільової функції. Вибирають тих претендентів, для яких цільова функція є меншою за деяке наперед визначене порогове значення. Після застосування операторів відбору застосовують оператор рекомбінації, який полягає в обміні генами батьківських хромосом при утворенні нової хромосоми. При дискретній рекомбінації гени вибирають випаково із заданою функцією розподілу по батьківських генах. Під генами розуміють складові частини хромосоми, які є вхідними параметрами оптимізаційної задачі. У нашій задачі у ролі генів виступають індекси семантичних полів. Якщо задача полягає у мінімізації цільової функції, тоді із більшою ймовірністю вибирають ген тієї батьківської хромосоми, у якої це значення цільової функції є мінімальним. Оператор кросовера може бути одноточковим, N-точковим та розсіяним. У одноточковому кросовері у послідовності генів вибирають точку розриву. Вибір точки розриву визначають випадковим чином із заданою функцією розподілу. Далі обмінюють ділянки генів у батьківських генах. У N-точковому кросовері існує N точок поділу хромосоми на ділянки генів, якими обмінюються батьківські хромосоми. У результаті обміну ділянок двох батьківских хромосом утворюють дві дочірні хромосоми нової популяції. В операторі розсіяного або однорідного кросовера (scatered crossover, uniform crossover) використовують бінарний вектор, який відіграє роль маски обміну генами. Розподіл бінарних значень у такій масці визначають заданим розподілом, зокрема, вибирають рівномірний розподіл. Розмір такого бінарного вектора рівний розміру батьківських хромосом. У дочірну хромосому попадає ген першої батьківської хромосоми, якщо значення складової бінарного вектора-маски, яке знаходиться на тому ж порядковому місці, рівне «0». Якщо значення відповідної складової бінарного вектора рівне «1», тоді дочірній хромосомі надають відповідний цьому місцю ген другої батьківської хромосоми. Оператор мутації використовують для зміни окремих генів у новостворених хромосомах. Ці зміни можуть відбуватись в одній або декількох заданих точках хромосоми. Ймовірність мутації задають деякою функцією розподілу, яка визначена характером та умовами задачі. Використання оператора мутації зумовлене необхідністю виведення популяції із локального мінімуму цільової функції для задач з існуванням локальних та глобальних мінімумів. Для формування нової популяції використовують оператори відбору хромосом. При використанні відбору відсіканням формують нову популяцію із батьківських та дочірних хромосом, які випадково відбирають із ймовірністю, що визначається значенням цільової функції. Причому у відборі беруть участь ті хромосоми, в яких значення цільової функції менше за визначений поріг. Вибір здійснюють до тих пір, поки не отримають нову популяцію із такою ж

кількістю хромосом, як і у попередній популяції. Очевидно, що деякі хромосоми можуть увійти у нову популяцію декілька разів. Також визначають деяку кількість хромосом із значенням цільової функції менше порогу, які можуть увійти у нову популяцію. При елітарному відборі задають процент батьківських та дочірних хромосом із найвищим значенням цільової функції, які увійдуть у нову популяцію без генетичних змін. При використанні такого підходу у кожній популяції буде знаходитись сукупність елітарних хромосом, які є найкращі на заданий момент розв'язку. Коли будуть знайдені кращі хромосоми у наступних популяціях, тоді вже вони стануть елітарними, а попередні елітарні хромосоми стануть звичайними. Часто різні методи відбору хромосом об'єднують у комбінований оператор відбору. У деяких алгоритмах використовують правило репродукції Холанда, за яким хромосоми із значенням цільової функції вище середнього значення копіюють у наступну популяцію, а хромосоми із значенням цільової функції, яке є меншим за середнє значення, видаляють [11]. Класичний генетичний алгоритм містить такі кроки:

1. Утворюють початкову популяцію із n хромосом.

2. Для кожної хромосоми визначають цільову функцію.

3. На основі заданого правила відбору вибирають дві батьківські хромосоми, на основі яких буде утворена нова дочірня хромосома для наступної популяції.

4. До відібраних батьківських пар застосовують оператор кросовера, за допомогою якого утворюють нову дочірню хромосому.

5. Здійснюють мутацію нащадків із деякою заданою ймовірністю.

6. Повторюють кроки 3–5, доки не буде згенерована нова популяція n хромосом.

7. Кроки 2–6 повторюють до тих пір, поки не будуть виконуватись умови зупинки алгоритму. Такою умовою може бути, наприклад, задане значення цільової функції або максимальна кількість ітерацій.

У дискретній оптимізації за допомогою генетичних алгоритмів кількість кроків, необхідних для пошуку оптимальних наборів вхідних параметрів, є поліноміально меншою у порівнянні із перебором можливих варіантів. Це пов'язано із наявністю деяких ділянок у хромосомах, які чимось подібні поведінкою на гени і які сукупно вносять оптимізаційний вклад у цільову функцію. Тобто вхідні параметри розглядають деякими групами (генами), якими обмінюються хромосоми за допомогою оператора кросовера, що суттєво зменшує кількість комбінацій параметрів в оптимізаційному аналізі.

5. Теоретико-множинна модель генетичного відбору семантичних полів

Розглянемо теоретико-множинну модель генетичного алгоритму оптимізації відбору семантичних полів для утворення семантичного простору текстових документів. Еволюцію генетичної оптимізації розглянемо у вигляді впорядкованої множини популяцій:

$$Ev^s = \left\{ Pop_k^s \mid k = 1,2,... \mid Ev^s \mid \right\}. \quad (1)$$

Вважаємо, що одне покоління хромосом утворюється однією популяцією. Популяція складається із множини хромосом:

$$Pop_k^s = \left\{ \chi_{jk}^{sp} \mid j = 1,2,... \mid Pop_k^{\ s} \mid ; k = 1,2,... \mid Ev^s \mid \right\}. \quad (2)$$

У загальному випадку різні популяції можуть містити різну кількість хромосом. У спрощеному випадку вважаємо, що кількість хромосом є однакова в усіх популяціях, тобто:

$$\mid Pop_k^{\ s} \mid = \mid Pop^{\ s} \mid = N_{pop}^{\chi}. \quad (3)$$

Кожну хромосому розглянемо як набір семантичних полів:

$$\chi_{jk}^{sp} = \left\{ s_{ijk}^{fxp} \mid i = 1,2... \mid \chi^s \mid ; j = 1,2,... \mid Pop_k{}^s \mid ; k = 1,2,... \mid Ev^s \mid \right\}, \quad (4)$$

де верхні індекси s_{ijk}^{fxp} позначають назви нижніх індексів: f – індекс семантичного поля, χ – індекс хромосоми, p – індекс популяції. Оператор одноточкового кросинговера розглянемо у вигляді

$$Crossover^p(\chi_{1k}^{sp}, \chi_{2k}^{sp}, m): \begin{cases} s_{11k}^{fxp} s_{21k}^{fxp} ... s_{m1k}^{fxp} ... s_{|\chi^s|1k}^{fxp} \\ s_{12k}^{fxp} s_{22k}^{fxp} ... s_{m2k}^{fxp} ... s_{|\chi^s|2k}^{fxp} \end{cases} \Rightarrow \begin{cases} s_{11k}^{fxp} s_{21k}^{fxp} ... s_{m2k}^{fxp} ... s_{|\chi^s|2k}^{fxp} \\ s_{12k}^{fxp} s_{22k}^{fxp} ... s_{m1k}^{fxp} ... s_{|\chi^s|1k}^{fxp} \end{cases}. \quad (5)$$

Індекс m позначає точку поділу хромосоми на дві частини семантичних полів, які обмінюють у батьківських хромосомах для утворення двох дочірніх хромосом наступної популяції. Оператор мутації розглянемо у вигляді

$$Mutation(\chi_{jk}^{sp}, m): s_{1jk}^{fxp} s_{2jk}^{fxp} ... s_{mjk}^{fxp} ... s_{|\chi^s|jk}^{fxp} \Rightarrow s_{1jk}^{fxp} s_{2jk}^{fxp} ... \tilde{s}_{mjk}^{fxp} ... s_{|\chi^s|jk}^{fxp}. \quad (6)$$

Внаслідок дії оператора $Mutation(\chi_{jk}^{sp}, m)$ змінюється семантичне поле s_{mjk}^{fxp} на семантичне поле $\tilde{s}_{mjk}^{fxp}$. Текстові документи представимо у семантичному просторі у вигляді вектора текстових частот семантичних полів p_{kj}^{sd}, які відображають частоту семантичного поля S_k у текстовому документі d_j. Значення частот p_{kj}^{sd} визначені як суми текстових частот лексем в аналізованому документі d_j, які належать заданому семантичному полю S_k. Сукупність значень p_{kj}^{sd} утворюють матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd} \right)_{k=1,\, j=1}^{N_s, N_d}. \quad (7)$$

Формування матриці частоти_семантичних_полів-документи розглянуто у роботах [1–3]. Значення частот p_{kj}^{sd} визначені як суми текстових частот в аналізованому документі, які належать заданому семантичному полю. Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd} \right)^T \quad (8)$$

відображає документ d_j в N_s-мірному просторі текстових документів із базисом, утвореним семантичними полями. Розглянемо використання генетичного алгоритму для оптимізації набору семантичних полів у задачі класифікації текстових документів. Як цільову функцію для еволюційної оптимізації набору семантичних полів базису семантичного простору розглянемо точність класифікатора. 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 експериментальному аналізі такі категорії утворюють групи новин. Множину цих категорій позначимо:

$$Categories = \{ Ctg_m \mid m = 1,2,..., N_{ctg} \}, \quad (9)$$

де $N_{ctg} = |Categories|$ визначає розмір множини категорій. За даними категоріями розподілені текстові документи множини D. Завдання полягає у пошуку цільової функції, яку описує відображення

$$F_{d\to ctg} : Categories \times D \to \{0,1\}. \quad (10)$$

Для характеристики класифікаторів використовують поняття точності (precision) [13, 14]. Точність класифікатора для категорії Ctg_j визначають як відношення кількості елементів, які правильно класифіковані як належні до категорії Ctg_j до загальної кількості елементів, які класифіковані як належні до категорії Ctg_j:

$$\Pr_j = \frac{\left|\left\{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\right\}\right|}{\left|\left\{d_i \mid Class(d_i) = Ctg_j\right\}\right|}, \quad (11)$$

де $Class(d_i)$ – визначена класифікатором категорія. Цільову функцію генетичної оптимізації визначимо так:

$$F_s^{ga} = 1 - \Pr_{avg}, \quad (12)$$

де $\Pr_{avg}$ – усереднена за всіма категоріями точність класифікатора. Завдання генетичної оптимізації буде полягати у мінімізації цільової функції F_s^{ga}. Існують підходи, які базуються на мінімізації інших типів функцій. У роботі [12] розглянуто генетичні алгоритми, які базуються на мінімізації штрафних функцій. Штрафні функції обраховують на основі цільових функцій із урахуванням обмежень на мінімальні та максимальні значення вхідних параметрів. Такий підхід часто використовують при оптимізації цілочисельних параметрів.

Як метод класифікації у дослідженні генетичної оптимізації розглянемо класифікацію за найближчими k сусідами, яку називають kNN класифікацією [13–15]. Цей метод відносять до векторних класифікаторів. В основі векторних методів класифікації лежить гіпотеза компактності. Згідно із цією гіпотезою, документи, які належать одному і тому ж классу, утворюють компактну область, а області, які належать різним классам, не перетинаються. Як міру близькості між документами виберемо евклідову відстань. У kNN класифікації границі категорій визначають локально. Деякий документ відносять 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. У генетичному відборі семантичних полів як вхідні параметри задачі оптимізації використаємо індекси масиву семантичних полів. Результатом генетичної оптимізації буде масив індексів, який визначає оптимальний набір семантичних полів.

6. Експериментальні дослідження

Для експериментального вивчення класифікації текстових документів у просторі семантичних полів ми вибрали стандартизовану текстову базу повідомлень груп новин 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 поля 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 лек-

сем. У семантичні поля також увійшли 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} типу частота_семантичного_поля–документ. Навчальну та тестову вибірки було вибрано рівними загальному об'єму аналізованого текстового масиву повідомлень. Для обчислень були використані генетичні алгоритми пакета прикладних програм Matlab. Для реалізації генетичної оптимізації складу семантичних полів у класифікаційному аналізі використано оператор розсіяного кросовера, однорідну селекцію батьківських хромосом та наявність групи елітарних хромосом. Оптимальні значення були вибрані експериментальним шляхом. Аналіз проведено при різних значеннях параметрів оптимізації. Розглянуто популяції розміром 30 хромосом. Кількість елітарних хромосом рівна 3. На рис. 1–3 наведена динаміка мінімального $F_{s(\min)}^{ga}$ та середнього значення $F_{s(avg)}^{ga}$ цільової функції F_s^{ga} із різними значеннями частки хромосом, утворених оператором кросовера та оператором мутації. На рис. 1 частка кросовера рівна 1, на рис. 2 – 0,8, на рис. 3 – 0,5. Загальна кількість аналізованих семантичних полів рівна 41. Розмір семантичних хромосом був вибраний рівним 5. Тобто, здійснювалась генетична оптимізація набору із 5 семантичних полів, для яких класифікація здійснювалась із найменшою похибкою. На рис. 1 спостерігається динаміка швидкого спадання середнього значення цільової функції до мінімального значення. Це зумовлено відсутністю у популяції нових значень вхідних параметрів. Генетичний відбір здійснюється лише на основі тих вхідних індексів семантичних полів, які знаходились у початковій популяції і були випадковим чином згенеровані. На рис. 2 середнє значення цільової функції наближується до мінімального значення і коливається в околі деякого значення. Ці коливання зумовлені наявністю хромосом, утворених за допомогою оператора мутації. Мутації дають можливість отримувати хромосоми із новими значеннями вхідних параметрів, що є ефективним у випадку наявності локальних мінімумів цільової функції. Поява нових значень індексів семантичних полів дає можливість генетичному алгоритму вийти із області можливого локального мінімуму і продовжити пошук глобального мінімуму цільової функції. Динаміка середнього значення цільової функції на рис.3 характеризується коливаннями на значній відстані від мінімального значення, що зумовлено малою фракцією хромосом, утворених оператором кросовера. Як випливає із отриманих даних, підбір параметрів генетичної оптимізації, зокрема, частки кросовера, є важливим для ефективного пошуку глобального мінімуму цільової функції. В аналізованих дослідженнях оптимальне значення кросовера рівне 0,8. Також досліджувався вплив елітарних хромосом. При зменшенні кількості елітарних хромосом із 3 до 1 суттєво збільшувався розкид середніх значень цільової функції у послідовних популяціях. У результаті генетичної оптимізації отримано мінімальне значення цільової функції, рівне 0,1023. Це значення відповідає набору індексів, які визначають такі семантичні поля у класифікації WordNet: noun.event, noun.phenomenon, verb.competition, verb.possession, verb.weather. Для отриманого оптимізованого набору семантичних полів проведено розрахунок точності класифікатора за найближчими сусідами для різних груп новин. Результати цього розрахунку наведені на рис. 4. Для різних груп новин характерна різна точність класифікації. Поряд із точністю Pr_j для різних категорій новин на рис. 4 наведено повноту (recall) класифікатора Rc_j, яку визначають як відношення успішно класифікованих документів у заданій категорії до загальної кількості документів у цій категорії. Ця характеристика є доповнюючою характеристикою до точності класифікатора.

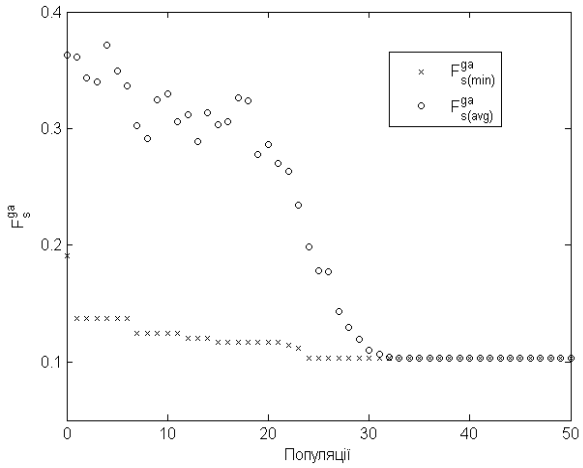

Рис. 1. Динаміка мінімального та середнього значень цільової функції при фракції кросовера, рівній 1

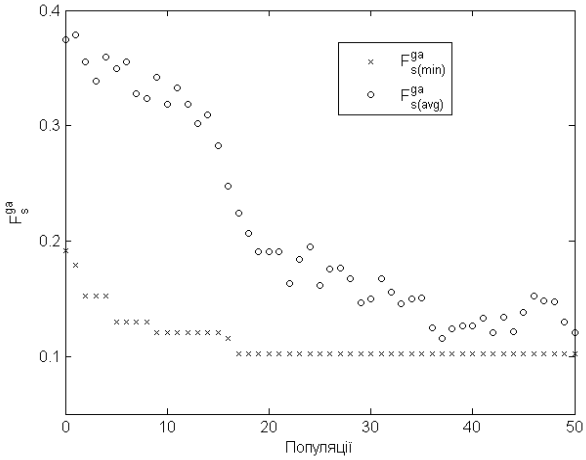

Рис. 2. Динаміка мінімального та середнього значень цільової функції при фракції кросовера, рівній 0,8

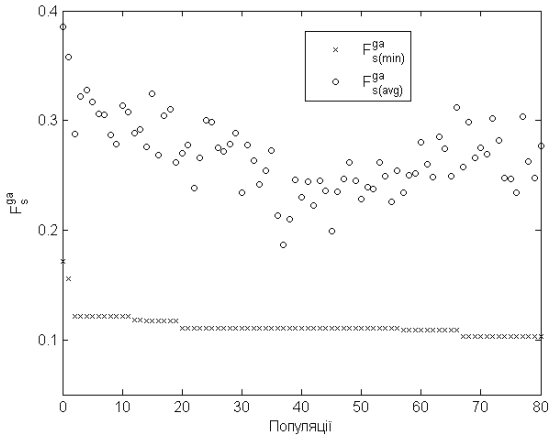

Рис. 3. Динаміка мінімального та середнього значень цільової функції при фракції кросовера, рівній 0,5

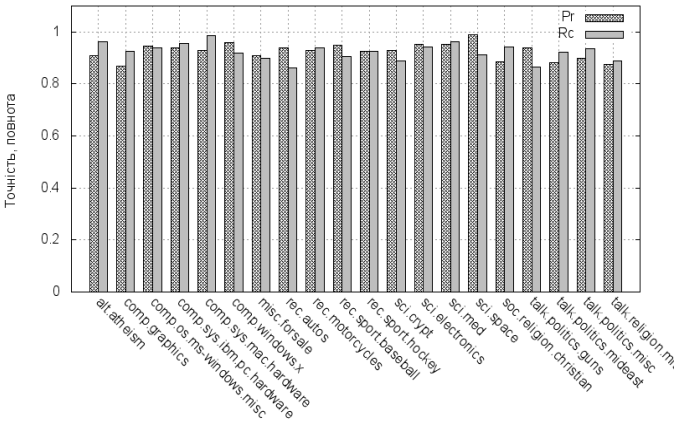

Рис. 4. Точність та повнота класифікатора за найближчими k сусідами (k=3) для груп новин в оптимізованому базисі семантичних полів

7. Висновки

Використання базису семантичних полів у моделі векторного простору текстових документів є ефективним для класифікаційного аналізу текстових даних. За допомогою генетичних алгоритмів можна оптимізувати набір семантичних полів, які утворюють векторний простір документів в алгоритмах інтелектуального аналізу текстових даних. Як цільову функцію для генетичної оптимізації використано точність класифікатора за найближчими k сусідами. Проведені експериментальні дослідження на тестовій вибірці текстових повідомлень груп новин показують ефективність використання генетичних алгоритмів для оптимізації набору семантичних полів, які утворюють базис векторного простору документів у класифікаційному аналізі текстових документів.

СПИСОК ЛІТЕРАТУРИ

1. Павлишенко Б.М. Використання концепції семантичного поля у векторній моделі текстових документів / Б.М. Павлишенко // 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 / Б.М. Павлишенко // 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 / Б.М. Павлишенко // Математичні машини і системи. – 2012. – № 1. – С. 69 – 76.
4. Pantel P. From Frequency to Meaning: Vector Space Models of Semantics/ P. Pantel, P.D. Turney // Journal of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141 – 188.
5. Fellbaum C. WordNet. An Electronic Lexical Database / C. Fellbaum. – Cambridge, MA: MIT Press. – 1998. – 432 p.
6. Gliozzo A. Semantic Domains in Computational Linguistics / A. Gliozzo, C. Strapparava. – Springer, 2009. – 132 p.
7. Goldberg D.E. Genetic Algorithms and Machine Learning / D.E. Goldberg, J.H. Holland // Machine Learning. – 1988. – Vol. 3, N 2–3. – P. 95 – 99.
8. Booker L.B. Classifier Systems and Genetic Algorithms / L.B. Booker, D.E. Goldberg, J.H. Holland // Artificial Intelligence. – 1989. – Vol. 40, N 1–3. – P. 235 – 282.
9. Батищев Д.И. Применение генетических алгоритмов к решению задач дискретной оптимизации / Батищев Д.И., Неймарк Е.А., Старостин Н.В. – Нижний Новгород, 2007. – 85 с.
10. Панченко Т.В. Генетические алгоритмы / Панченко Т.В.; под ред. Ю.Ю. Тарасевича. – Астрахань: Астраханский университет, 2007. – 87 с.
11. Гладков Л.А. Генетические алгоритмы / Гладков Л.А., Курейчик В.В., Курейчик В.М.; под ред. В.М. Курейчика. – [2-е изд.]. – М.: ФИЗМАТЛИТ, 2006. – 320 с.
12. Deb K. An efficient constraint handling method for genetic algorithms / K. Deb // Computer Methods in Applied Mechanics and Engineering. – 2000. – P. 311 – 338.
13. Manning C.D. Introduction to Information Retrieval / C.D. Manning, P. Raghavan, H. Schütze. – Cambridge University Press, 2008. – 496 p.
14. Sebastiani F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys. – 2002. – Vol. 34, N 1. – P. 1 – 47.
15. Анализ данных и процессов: учеб. пособ. / А.А. Брасегян, М.С. Куприянов, И.И. Холод [и др.]. – СПб.: БХВ-Петербург, 2009. – 512 с.

Стаття надійшла до редакції 09.11.2012