УДК 004.825

Волосюк Ю. В., канд. техн. наук (Тел.: +380 98 438 75 40. E-mail: relax_eu@mail.ru)
(Миколаївський філіал ПВНЗ «Європейський Університет», м. Миколаїв)

МЕТОДИ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ В ЗАДАЧАХ TEXT MINING

Волосюк Ю. В. Методи класифікації текстових документів в задачах Text Mining. В роботі надана загальна постанова процедури класифікації текстів. Наведено огляд існуючих підходів до вирішення задачі класифікації. Описані основні підходи, що використовуються в задачі класифікації текстів, визначено етапи процесу класифікації та розглянуті найбільш поширені математичні методи класифікації текстових документів. Розкрито особливості використання, переваги та недоліки зазначених методів. Зроблено висновок щодо необхідності подальшого розроблення алгоритмів класифікації на основі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати при навчанні та високу якість класифікації в реальних завданнях.

***Ключові слова*:** класифікація, рубрикація, метод опорних векторів, дерева рішень, штучні нейроні мережі

Волосюк Ю. В. Методы классификации текстовых документов в задачах Text Mining. В работе представлена постановка общей процедуры классификации текстов. Приведен обзор существующих подходов к решению задачи классификации. Описаны основные подходы, используемые в задаче классификации текстов, определены этапы процесса классификации и рассмотрены наиболее распространенные математические методы классификации текстовых документов. Раскрыты особенности использования, преимущества и недостатки указанных методов. Сделан вывод о необходимости дальнейшей разработки алгоритмов классификации на основе указанных методов, которые были бы простыми в реализации, эффективными, имели небольшие вычислительные затраты при обучении и высокое качество классификации в реальных задачах.

***Ключевые слова*:** классификация, рубрикация, метод опорных векторов, деревья решений, искусственные нейронные сети

Volosyuk Yu. V. Discriminatory analyses of text documents in problems Text Mining. In work, the total characteristic of procedure of classification of texts is presented. The review of existing approaches is led to a solution of a problem of classification. The basic approaches used in a problem of classification of texts are described, stages of process of classification are defined and the most widespread mathematical discriminatory analyses of text documents are considered. Singularities of use, advantage and shortages of the specified methods are uncovered. The conclusion is drawn on necessity of the further working out of algorithms of classification based on the specified methods, which would be simple in realization, effective, had small computing expenditures at training and high quality of classification in real problems.

***Keywords*:** classification, a rubricating, a method of basic vectors, trees of solutions, artificial neural network

Вступ і постановка проблеми. Класифікація або рубрикація інформації (віднесення порції інформації до однієї або декількох категорій з обмеженої множини) є традиційним завданням організації знань і обміну інформацією.

Запропоновано багато методів для вирішення даного завдання за допомогою автоматичних процедур. Основне призначення даних методів – аналіз, класифікація та виявлення прихованих закономірностей у великих обсягах різнорідних складно структурованих даних. Існуючі методи доцільно поділити на два принципово різних класи: методи машинного навчання і методи, засновані на знаннях (так званий "інженерний підхід"). При використанні методів машинного навчання для побудови класифікатора використовується колекція документів, заздалегідь відрубрикована людиною. Алгоритм машинного навчання будує процедуру класифікації документів на основі автоматичного аналізу заданої множини відрубрикованих текстів. При використанні методів, заснованих на знаннях, правила віднесення документа до тієї або іншої рубрики задаються експертами на основі аналізу рубрикатора і, можливо, частини текстів, що належать рубрикуванню.

Незважаючи на велику кількість запропонованих методів класифікації текстів, питання обрання такого методу, який був би простим в реалізації, ефективним, мав низькі

обчислювальні витрати при навчанні та класифікації і високу якість класифікації в більшості реальних завдань, є актуальним завданням.

Метою статті є опис основних підходів, що використовуються в задачі класифікації текстів, аналіз математичних методів класифікації текстових документів, визначення особливостей використання, переваг та недоліків зазначених методик.

Терміни і визначення. Під класифікацією текстів (*Text Categorization*) розуміється розподіл текстових документів по заздалегідь визначених категоріях. Методи класифікації текстів лежать на межі двох областей – машинного навчання (*machine learning*) і інформаційного пошуку (*information retrieval*) [1]. Відповідно, автоматична класифікація може здійснюватися на основі заздалегідь заданої схеми класифікації і вже наявної множини класифікованих документів або бути повністю автоматизованою. При використанні підходів машинного навчання, класифікаційне правило будується на основі тренувальної колекції текстів.

Процес класифікації складається з двох етапів: конструювання моделі і її використання. Конструювання моделі – це опис множини класів (отримана модель може бути представлена класифікаційними правилами, деревом рішень, математичною формулою). Використання моделі полягає в класифікації нових або невідомих значень, оцінка точності моделі.

Формально задачу класифікації можливо описати таким чином: передбачається, що алгоритм класифікації працює на деякій множині документів $D = \{d_i\}$. Вся множина документів розбивається на непересічні підмножини класів:

$$C = \{Ci\}, \bigcup_{d \in Ci}^{\forall i} d = D, Ci \cap Cj = 0 (i \neq j).$$

Завданням класифікації є визначення класу, до якого відноситься даний документ. Кожному елементу d ставиться у відповідність набір ознак $d = \{X_i\}$. Далі застосовується алгоритм класифікації для виділення документів найбільш відповідних заданому класу [2].

Для класифікації застосовуються різноманітні методи, кожен з яких має свої переваги і особливості використання. Переважна більшість методів класифікації текстів так чи інакше засновані на припущені, що документи, які відносяться до однієї категорії, мають однакові ознаки (слова чи словосполучення), і наявність чи відсутність таких ознак в документі визначає його приналежність чи неприналежність до тієї чи іншої теми. Таким чином, для кожної категорії повинна бути множина ознак:

$$F(C) = \cup(c_r),$$

де $F(c_r) =< f_1, ..., f_k, ..., f_z >$. Таку множину ознак називають словником, через те, що вона складається з лексем, які включають слова і/чи словосполучення, що характеризують категорію. Подібно категоріям, кожен документ також має ознаки, по яким його можливо віднести з деяким ступенем вірогідності до однієї чи декільком категоріям:

$$F(d) =< f_1^i, ..., f_l^i, ..., f_y^i >.$$

Множина ознак усіх документів повинна співпадати з множиною ознак категорій, тобто:

$$F(C) = F(D) = \cup F(d_i).$$

Необхідно зазначити, що вказаний набір ознак є відмінною рисою класифікації документів порівняно з класифікацією об'єктів в Data Mining, які характеризуються набором атрибутів. Прийняття рішення щодо відношення документа d_i до категорії c_r відбувається на основі перетину:

$$F(d_i) \cup F(c_r).$$

Задача методів класифікації текстів полягає в тому, щоб якнайкраще обрати такі ознаки і сформулювати правила, опираючись на які буде прийматися рішення щодо віднесення документа до рубрики. Визначимо й опишемо основні з цих методів, до яких відносяться:

- класифікація за допомогою дерев рішень;
- Байесівська (наївна) класифікація;
- класифікація методом опорних векторів;
- класифікація за допомогою методу найближчого сусіда;
- класифікація за допомогою штучних нейронних мереж.

Класифікація за допомогою дерева рішень. Дерева рішень (decision trees) розбивають дані на групи на основі значень змінних простору ознак, внаслідок чого виникає ієрархія операторів "якщо-то", які класифікують дані [3]. Для прийняття рішення, до якої категорії віднести даний документ, потрібно відповісти на питання, що знаходяться у вузлах цього дерева, починаючи з його кореня. Питання мають вигляд – "значення змінної x_i більше порога b_i?". Якщо відповідь позитивна, здійснюється перехід до правого вузла цього дерева, якщо негативна – до лівого вузла. Наступне питання пов'язане з відповідним вузлом.

Для автоматичної побудови дерев рішень за допомогою навчання на прикладах розроблено ряд алгоритмів [4]. Розглянемо один з таких алгоритмів – CLS. Цей алгоритм циклічно розбиває навчальні приклади на класи відповідно до змінної, що має найбільшу класифікуючу силу. Кожна підмножина прикладів, що виділяється такою змінної, знову розбивається на класи з використанням змінної з найбільшою класифікуючою здатністю і т. д. Розбиття закінчується тоді, коли в підмножині виявляються лише елементи з одного класу. В ході процесу утворюється дерево рішень.

Для визначення змінної з найбільшою класифікуючою силою використовується критерій інформаційної ваги слова в рубриці. Зазвичай після побудови "точного" дерева рішень до отриманого дерева застосовуються різні процедури усічення і перетворення дерева для того, щоб забезпечити баланс між складністю дерева (кількістю вузлів) і якістю навчання. Класичним підходом для перетворення дерев рішень є алгоритм C4.5.

Одним з головних недоліків методу дерев рішень для задач класифікації текстів є той факт, що алгоритм побудови дерев рішень надає однакову вагу "позитивним" і "негативним" розгалуженням у вузлах. Велика кількість "негативних" гілок в описі рубрики може призводити до важко інтерпретуємих правил і "перенавчанню" алгоритму класифікації.

До переваг зазначеного методу слід віднести той факт, що побудоване дерево легко піддається аналізу, результат роботи алгоритму можна інтерпретувати в наочних термінах. Існують програми наочного графічного відображення дерев рішень.

Метод Байєсової (наївної) класифікації (Naive Bayes). Наївний Байєсівський класифікатор традиційно використовується в задачах класифікації текстів, таких як фільтрація спаму, автоматична рубрикація або визначення тональності документа. Набули поширеного розвитку два його різновиди: багатомірний (multivariate) та мультиноміальний (multinomial).

У загальному вигляді визначення найбільш вірогідного класу алгоритмом наївною байєсівської класифікації виглядає наступним чином:

$$C = \arg\max_{c \in C} P(C \mid o_1, o_2, \ldots o_n) = \arg\max_{c \in C} P(c) \prod P(o_i \mid c),$$

де C – набір класів, а $o_1, o_2, \ldots o_n$ – набір ознак. Класифікація зводиться до обчислення максимального значення аргументу, при відомому наборі незалежних ознак $o_1, o_2, \ldots o_n$.

При цьому:

$$P(c)\prod P(o_i \mid c) = P(C)P(P(o_1 \mid c)P(o_2 \mid c)...P(o_n \mid c).$$

Обчислення ймовірності класу $P(C)$ при відомих ознаках $o_1, o_2, ...o_n$ зводиться до наступного:

$$P(C \mid o_1\, o_2 \,...o_n) = \sum(o_1\, o_2 \,...o_n) + 1 / \sum(\, C \mid A) + \sum A,$$

де A – набір відомих ознак, отриманих при навчанні класифікатора.

Класифікація тексту при цьому виглядає наступним чином:

$$C(T) = \max \sum(t_1, t_2 ... t_{n1} \mid C),$$

де T – текст, що класифікується, а $t_1, t_2 ... t_{n1}$ – набір речень тексту. Так, приналежність тексту до того чи іншого класу зводиться до обчислення максимального значення суми коефіцієнтів приналежності реченням.

Зазначений метод використовує ймовірнісну модель, в якій класифікація та включення у відповідну категорію документів проводиться шляхом оцінювання ймовірності появи слів у документі. Ймовірності можуть бути використані для оцінки найбільш близьких категорій тестового документа [6].

Основні переваги наївного байєсівського класифікатора – простота реалізації і низькі обчислювальні витрати при навчанні та класифікації. У тих рідкісних випадках, коли ознаки дійсно незалежні (або майже незалежні), наївний байєсівський класифікатор (майже) оптимальний. Основним недоліком методу є відносно невисока якість класифікації в більшості реальних завдань. Зазначений метод часто використовується в якості базового методу при порівнянні різних методів машинного навчання.

Метод опорних векторів (Support Vector Machine, SVM) – використовує процес пошуку площини вирішення, яка може розділити позитивні і негативні приклади в багатовимірному просторі функції, в якому навчальні документи представлені як вектори. Цей метод розроблений В. Вапником в 1995 році, був вперше застосований до задачі класифікації текстів Торстеном Джохімсом. У своєму первинному вигляді алгоритм вирішував завдання розрізнення об'єктів двох класів. Метод набув величезну популярність завдяки своїй високій ефективності. Багато дослідників використовують його в роботах, присвячених класифікації текстів. Підхід, запропонований В. Вапником для визначення того, до якого з двох заздалегідь визначених класів повинен належати аналізований зразок, заснований на принципі структурної мінімізації ризику.

Результати класифікації текстів за допомогою методу опорних векторів є одними з найкращих, у порівнянні з іншими методами машинного навчання [9]. Однак, швидкість навчання даного алгоритму одна з найнижчих. Метод опорних векторів вимагає великого обсягу пам'яті і значних витрат машинного часу на навчання, що знижує його масштабованість. Проте даний алгоритм можна використовувати як еталон з точки зору якості класифікації [10], так метод буде працювати ефективно, якщо опорних векторів буде порівняно не багато, якщо ж їх кількість зростає, метод стає малоефективний через значно збільшену складність.

Метод *k*-найближчого сусіда (*k*-nearest neighbor). Цей метод є одним з найбільш вивчених і високоточних алгоритмів, що використовуються при створенні автоматичних класифікаторів. Вперше він був запропонований ще в 1952 році для вирішення завдань дискримінантного аналізу. У дослідженнях, присвячених аналізу роботи різних алгоритмів

машинного навчання для задачі класифікації текстів, цей метод демонструє одні з найкращих результатів [11].

В основі методу лежить досить проста ідея: знаходити та відрубриковувати колекції найбільш схожі на аналізуємий текст документи і, опираючись на знання про їх категоріальну приналежність, класифікувати невідомий документ. З метою визначення рубрики, релевантні документу *d*, цей документ порівнюється з усіма документами з навчальної вибірки. Для кожного документу *e* з навчальної вибірки знаходиться вибірка – косинус кута між векторами ознак:

$$\rho(d,e) = \cos(d,e)$$

Надалі з навчальної вибірки обираються *k* документів, найближчих до *d* (*k*-параметри). Для кожної рубрики обчислюється релевантність за формулою:

$$s(c_j,d) = \sum_{e\in\{k \text{ найближчих сусідів}\}\wedge c_j\in Rub(e)} \cos(d,e)$$

Рубрики, що мають релевантність вище деякого заданого порогу, вважаються відповідними документу. Параметр *k* зазвичай обирається в інтервалі від 0 до 100. При монотематичній категоризації обирається рубрика з максимальним значенням. Якщо ж документ може бути приписаний до декількох рубрик (випадок мультитематичної категоризації), класи вважаються відповідними, якщо значення перевищує деякий наперед заданий поріг.

Головною особливістю, що виділяє зазначений метод серед інших, є відсутність у нього стадії навчання. Іншими словами, належність документа рубрикам визначається без побудови класифікуючої функції. Основною перевагою такого підходу є можливість оновлювати навчальну вибірку без перенавчання класифікатора. Ця властивість може бути корисною, наприклад, у випадках, коли навчальна колекція часто поповнюється новими документами, а перенавчання займає надто багато часу. Класичний алгоритм пропонує порівнювати аналізований документ з усіма документами з навчальної вибірки і тому головний недолік описаного методу полягає в тривалості часу роботи рубрикатора на етапі класифікації [12].

Штучні нейронні мережі (Artificial Neural Network). Штучні нейронні мережі набули широкого вивчення в галузі штучного інтелекту для аналізу даних з 1986 року. Вони представляють собою математичну модель, а також її програмні або апаратні реалізації, побудовані за подібністю мереж нервових клітин живого організму. Нейронні мережі – це один з найбільш відомих і старих методів машинного навчання.

Штучні нейронні мережі – це адаптивна система, яка складається з групи з'єднаних штучних нейронів. Система може бути навчена для зміни її внутрішніх станів, відображення зв'язків документів та їх категорій. Для ефективного проведення класифікації текстів необхідно визначити раціональну структуру і топологію нейронної мережі. Основні топології класифікуючих нейронних мереж – це одно- і багатошаровий персептрон, нейромережевий Гаусів класифікатор, мережа Кохонена, мережа вбудованого розповсюдження, каскадна мережа [13]. Всі вищевказані топології мають високу точність в обробці одночасно лінійних і нелінійних прикладів, але прийняття рішень щодо класифікації важко формалізуються у зв'язку з природою організації нейронної мережі та представляють нетривіальну задачу з урахуванням масштабованості з обмеженими обчислювальними ресурсами.

Висновки з даного дослідження і перспективи подальшого розвитку у даному напрямку. В процесі дослідження методів класифікації текстових документів надано загальну процедуру класифікації текстів, наведено огляд існуючих підходів до вирішення

задачі класифікації, описані основні підходи, що використовуються в задачі класифікації текстів, визначено етапи процесу класифікації та розглянуті найбільш поширені математичні методи класифікації текстових документів. Розкриті особливості використання, переваги та недоліки зазначених методів дозволяють зробити висновок щодо необхідності подальшого вдосконалення алгоритмів класифікації на основі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати при навчанні та високу якість класифікації в реальних завданнях.

Література

1. Барсегян А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. – 3-е изд., перераб. и доп. – Санкт-Петербург : БХВ-Петербург, 2009. – 512 с.

2. Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99), 1999. – P. 42-49.

3. Вагин В. Н. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин, Е. Ю. Головина, А. А. Загорянская, М. В. Фомина. – Москва : Физматлит, 2004. – 704 с.

4. Quinlan J. R. C4.5 Programs for machine learning. – Morgan Kaufmann, – San Mateo, Californie, 1993.

5. Айвазян С. А. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – Москва : Финансы и статистика, 1989.

6. Joachims T. Making large-scale SVM learning practical / T. Joachims // Advances in Kernel Methods Support Vector Learning. – MIT Press, 1999. – 218 p.

7. Губин М. В. Модели и методы представления текстового документа в системах информационного поиска / М. В. Губин. – Санкт-Петербург : Санкт-Петербургский государственный университет, 2005.

8. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы: пер. с польк. / Д. Рутковская, М. Пилиньский, Л. М. Рутковский. – Москва : Горячая линия – Телеком, 2004. – 472 с.

9. Барсегян А. А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян. – Санкт-Петербург : БХВ-Петербург, 2007. – 384 с.

10. CJC Burges. A Tutorial on Support Vector Machines for Pattern Recognition [Електронний ресурс] // – Режим доступу : http://www.music.mcgill.ca/rfergu/adamTex/references/Burges98.pdf

11. Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. SIGIR'2012, 22nd ACM International Conference on Research and Development in Information Retrieval, 2012. – P. 42-49.

12. Sebastiani F. Machine learning in automated text categorization / F. Sebastiani // ACM Comput. Surv. – March 2010. – Vol. 34, No. 1. – P. 1-47.

13. Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99), 2007. – P. 42-49.