УДК 004.852

В. Ю. Бондарчук

ПОРІВНЯННЯ МЕТОДІВ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Обґрунтовано актуальність задачі визначення тональності текстів. Зазначено, що для розв'язання цієї задачі застосовують, зокрема наївний Байєсівський класифікатор метод максимальної ентропії, та метод опорних векторів. Для вибору потрібно керуватись, технічними можливостями та точністю результату і можна спробувати скомбінувати ці методи.

Ключові слова: аналіз тональності, наївний Байєсівський класифікатор, метод опорних векторів.

Abstract

The relevance of the task of determining the texts sentiment analysis is indicated. It is noted that in order to solve this problem, the naive Bayesian classifier, method of maximum entropy, and the method of reference vectors are used in particular. You need to be guided by the technical options and the result accuracy and you can try to combine these techniques.

Key words: sentiment analysis, naive Bayes classifier, support vector machine.

У сучасному світі, підприємства та організації завжди бажають дізнатися думки своїх споживачів або користувачів щодо власних продуктів чи послуг. В останні роки були проведені дослідження емоційного забарвлення відгуків про фільми, ресторани, готелі, техніку, політичні події, тощо [1]. Ці задачі дозволяє розв'язує так званий аналіз тональності тексту. Аналіз тональності тексту (англ. sentiment analysis) є відносно новим напрямком автоматизації аналізу емоційної складової тексту. Він набуває великої популярності у зв'язку з розвитком різних платформ для оцінювання (сайти про фільми, одяг, техніку тощо). Правильне його застосування дозволяє оцінити реакцію користувачів на той чи інший продукт і врахувати її у подальшому [2]. З метою розв'язання цих задач використовують, зокрема, наївний Байєсовський класифікатор, метод опорних векторів тощо [3].

Метод опорних векторів (Support Vector Machine, SVM) – передбачає процес пошуку площини рішення, яка може розділити позитивні та негативні приклади у багатовимірному просторі функції, в якому навчальні документи представлені як вектори. Цей метод, розроблений В. Вапником у 1995 році, був вперше застосований до задачі класифікації текстів Торстеном Джохімсом [4]. У своєму первинному вигляді алгоритм вирішував завдання розрізнення об'єктів двох класів. Метод набув величезну популярність завдяки високій ефективності. Багато дослідників використовують його у робо*т*ах, присвячених класифікації текстів. Метод, запропонований В. Вапником для визначення того, до якого з двох заздалегідь визначених класів повинен належати аналізований зразок, заснований на принципі структурної мінімізації ризику. SVM належить до розряду лінійних класифікаторів. Його перевагами є те, що він дозволяє отримати розв'язок, близький до оптимального, навіть без вбудованих знань про предметну область [5], при цьому, завдяки тому, що даний метод зводиться до розв'язання задачі квадратичного програмування на опуклому просторі – він гарантує єдиність розв'язку. Серед недоліків методу слід відзначити значне збільшення обчислювальної складності при збільшенні ефективності [6].

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

Зазначений метод використовує ймовірнісну модель, в якій класифікація та включення у відповідну категорію документів здійснюється шляхом оцінювання ймовірності появи слів у документі. Ймовірності можуть бути використані для оцінювання найбільш близьких категорій тестового документа [7].

Основні переваги наївного Байєсівського класифікатора – простота реалізації та низькі обчислювальні витрати під час навчання 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 у більшості реальних завдань. Зазначений метод часто використовується в якості базового під час порівняння різних методів машинного навчання.

Слід також згадати про класифікацію за допомогою методу максимальної ентропії. У випадку з розбивкою на два класи це використання логістичної регресії для пошуку розподілу даних за класами. На відміну від наївного Байєсівського класифікатора даний метод не передбачає незалежності ознак. Це означає, що можна використовувати для передбачення ознаки різної природи (наприклад, виміряти n-буквенні сполуки і словосполучення у повідомленні одночасно).

Також, у подальшому, можна було б спробувати використати гештальт-ранжування [8, 9] для прийняття рішень стосовно емоційного забарвлення текстів.

Висновки

У роботі обґрунтовано актуальність задачі аналізу тональності текстів. Вказано, що для розв'язання цієї задачі доцільно застосовувати метод опорних векторів, оскільки він є точнішим, але вимагає великої обчислювальної складності при збільшенні ефективності. З іншої сторони наївний Байєсівський класифікатор і метод максимальної ентропії є простими у реалізації, але і менш точними методами. Отже, для розв'язання вищевказаної задачі слід врахувати технічні можливості та точність результату і можна спробувати певним чином скомбінувати вищезазначені методи.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

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. Месюра В. І. Основи проектування систем штучного інтелекту. Навчальний посібник / В. І. Месюра, Л. М. Ваховська. – В.: ВДТУ, 2000. – 96 с.
4. 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.
5. Quinlan J. R. C4.5 Programs for machine learning. – Morgan Kaufmann, – San Mateo, Californie, 1993.
6. 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.
7. Joachims T. Making large-scale SVM learning practical / T. Joachims // Advances in Kernel Methods Support Vector Learning. – MIT Press, 1999. – 218 p.
8. В. Колодний, Д. Кудрявцев. Інформаційна технологія візуального моделювання та обробки тернарних гештальт-ранжувань, ІТКІ, vol 42, № 2, с. 26 – 34, жовтня 2018.
9. Застосування гештальт-ранжувань для виявлення переваг ОПР / В. В. Колодний, В. В. Зубко // «Інтернет-освіта-наука-2016» : Збірник матеріалів конференції. – Вінниця : ВНТУ, 2016. – С. 43 – 44.

Бондарчук Віталій Юрійович – студент гр. 2 КН-15б факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: 2kn15b.bondarchuk@gmail.com

Науковий керівник: ***Арсенюк Ігор Ростиславович*** – к. т. н., доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет

Vitalii Y. Bondarchuk – Student of Department of information technology and computer engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: 2kn15b.bondarchuk@gmail.com

Supervisor: ***Igor R. Arsenyuk*** – Cand. Sc. (Eng), Assistant Professor of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia