УДК 004.65 **DOI:** 10.31673/2412-9070.2019.065860

О. О. ПІДМОГИЛЬНИЙ, аспірант;
О. М. ТКАЧЕНКО, доктор техн. наук, доцент;
О. І. ГОЛУБЕНКО, ст. викладач;
О. В. ДРОБИК, канд. техн. наук, професор,
Державний університет телекомунікацій, Київ

NAIVE BAYES CLASSIFIER ЯК ОДИН ІЗ ВАРІАНТІВ ФІЛЬТРАЦІЇ НЕБАЖАНОЇ ЕЛЕКТРОННОЇ ПОШТИ

Розглянуто проблему класифікації електронної кореспонденції для визначення приналежності листів до небажаної електронної пошти. Запропоновано три підходи для розв'язання даної проблеми. Детально проаналізовано Naive Bayes Classifier як найбільш ефективний спосіб, що дає змогу з високою точністю класифікувати листи і водночас є економним вирішенням стосовно витрат обчислювальних ресурсів.

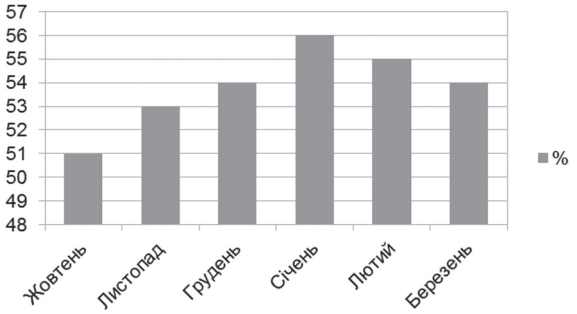
Ключові слова: Naive Bayes Classifier; обчислювальна ефективність; Rule based classification; Weight based classification; фільтрація спам-листів; класифікатор.

Вступ

За даними Лабораторії Касперського, у 2018-2019 роках поштовий спам порівняно зі світовим трафіком становив у середньому 53,8% (статистику за 6 місяців зображено на рисунку). Перші місця в рейтингу джерел спаму традиційно посідають такі країни, як Китай (15,82%), США (12,64%), Німеччина (5,86%), Росія (6,98%) та Бразилія (6,95%).

Учені з Каліфорнійського університету в Берклі підрахували, що на підтримання світової мережі Інтернет людство витрачає 2% від загального обсягу використання електроенергії цивілізації, тобто із споживаних 16 ТВт на опрацювання інтернет-трафіку припадає 0,32 ТВт енергії. Така статистика свідчить, що на оброблення небажаної електронної пошти витрачається майже 0,17 ТВт із середньою світовою ціною на електроенергію (2018–2019 роки) у 0,166 дол. за 1 кВт · год., отже, на спам-повідомлення витрачається 282 млрд дол.

З огляду на зазначену статистику і підрахунки можна дійти висновку щодо потреби економічно доцільного розроблення результативних методів боротьби зі спамом. Сьогодні існує кілька методів класифікації, які себе зарекомендували найбільш ефективно. Розглянемо деякі з них.



Статистика небажаної електронної пошти за 6 місяців

Основна частина

Rule based classification

Будь-яка класифікація ґрунтується на правилах (*rule based classification*). Ви імплементуєте правила визначення класу документа за його текстом у вигляді if-then-else виразів. Цей підхід може бути прийнятним варіантом, якщо ви працюєте з невеликою кількістю документів, котру здатні охопити і ретельно проаналізувати, оскільки чітко контролюєте правила, за якими класифікатор ухвалює рішення. Але цей підхід має очевидні мінуси — для того щоб вибрати значимі для класифікації слова, потрібно володіти експертними знаннями в предметній області. Чи є у вас, наприклад, розуміння з приводу ключових слів, які добре відрізняють документи фінансової тематики від документів економічної сфери? Вочевидь не завжди факт наявності або відсутності будь-якого одного слова є вирішальним чинником для ухвалення рішення.

Скажімо, якщо повернутися до завдання визначення спаму і трохи поміркувати про те, які слова є добрими класифікаційними ознаками (*classifying feature*), то можна зрозуміти, що немає жодного слова, наявність якого гарантувала б, що повідомлення є спамом.

Загалом, будь-яке з відомих спам-слів нехай і мало коли, але зустрічається в повсякденному житті. Тому ухвалювати остаточне рішення, орієнтуючись на факт наявності або відсутності будь-якого одного слова, ідея непродуктивна. Ми можемо ускладнювати правила, додаючи вкладені умови, але досить швидко зрозуміємо, що можливості людини в формулюванні таких правил дуже обмежені, оскільки складність правил зростає експоненціально з кількістю вибраних для класифікації слів.

Weight based classification

Класифікація на основі ваги кожного слова, яка визначатиме ймовірність, що повідомлення з цим словом є спамом (табл. 1).

Таблиця 1

Таблиця ваг

Слово	SPAM	NOSPAM
Показую	0,99	0,01
Заробити	0,40	0,60
Допомога	0,99	0,01
Собаці	0,01	0,99

Ми беремо кожне слово і розраховуємо сумарну вагу документа окремо для класу «спам» і класу «не-спам». Сумарна вага визначається як степінь ваг усіх відомих слів документа. Словами, для яких у нас немає ваги, у процесі класифікації можна знехтувати. Чия сумарна вага буде вищою, той клас і візьме гору. Це більш дієвий підхід, оскільки він гнучкіший і ухвалює рішення, ґрунтуючись на всі відомі слова в тексті.

Naive Bayes Classifier

Але нам потрібно щось краще і більш гнучке. Таке вирішення — це Naive Bayes Classifier, і для цього є низка причин:

- він простий в імплементації та тестуванні;
- процес навчання досить ефективний порівняно з іншими, більш складними класифікаторами;
- на невеликих корпусах документів різниця між іншими набагато складнішими алгоритмами класифікації часто істотна, а іноді Naive Bayes Classifier може бути і більш точним.

В основу Naive Bayes Classifier покладено теорему Байєса, яка дає змогу визначати ймовірність будь-якої події за умови, що сталася інша статистично взаємозалежна подія. Інакше кажучи, за формулою Байєса можна більш точно розрахувати ймовірність, узявши до уваги як раніше відому інформацію, так і дані нових спостережень. Формулу Байєса можна дістати з основних аксіом теорії ймовірностей, зокрема з умовної ймовірності. Особливість теореми Байєса полягає в тому, що для її практичного застосування потрібна велика кількість розрахунків, обчислень, тому байєсові оцінки стали активно використовувати тільки після революції в комп'ютерних та мережних технологіях:

$$P(c|d)=\frac{P(d|c)P(c)}{P(d)},$$

де: $P(c \mid d)$ — імовірність, що email d належить класу c, саме її треба розрахувати;

$P(d \mid c)$ — імовірність зустріти email d серед усіх листів класу c;

$P(c)$ — безумовна ймовірність зустріти лист класу c у корпусі листів;

$P(d)$ — безумовна ймовірність документа d у корпусі документів.

Отже, теорема Байєса дозволяє переставити місцями причину і наслідок. Знаючи з якою ймовірністю причина призводить до якоїсь події, ця теорема дає можливість розрахувати ймовірність того, що саме ця причина призвела до нинішньої події.

Оскільки мета класифікації полягає в розумінні до якого класу належить лист, тому потрібна не сама ймовірність, а найбільш імовірний клас. Байєсовий класифікатор використовує оцінку апостеріорного максимуму (*Maximum a posteriori estimation*) для визначення найбільш вірогідного класу.

Для реалізації байєсового класифікатора необхідна навчальна вибірка, в якій проставлено відповідності між текстовими листами та їхніми класами. Потім потрібно зібрати таку статистику з вибірки, яка використовуватиметься на етапі класифікації:

- відносні частоти класів у корпусі листів; тобто, як часто зустрічається лист того чи іншого класу;
- сумарна кількість слів у листах кожного класу;
- відносні частоти слів у межах кожного класу;
- розмір словника вибірки; кількість унікальних слів у вибірці.

Сукупність цієї інформації називатимемо моделлю класифікатора. Потім на етапі класифікації необхідно для кожного класу розрахувати значення наступного виразу і вибрати клас із максимальним значенням.

$$\log\left(\frac{Dc}{D}\right)+\sum_{i\in Q}\log\left(\frac{Wic+1}{|V|+Lc}\right),$$

де: Dc — кількість листів у навчальній вибірці, що належать класу c;

D — загальна кількість листів у навчальній вибірці;

$|V|$ — кількість унікальних слів у всіх листах навчальної вибірки;

Lc — сумарна кількість слів у листах класу c в навчальній вибірці;

Wic — скільки разів кожне слово зустрічалося в листах класу c в навчальній вибірці;

Q — множина слів листа, який перевіряється (зокрема повтори).

Нехай у нас є три листи, для яких відомі їхні класи:

[SPAM] показую, як заробити;
[SPAM] допомога собаці Чапі;
[NOSPAM] привіт, як справи.

Модель класифікатора наведено в табл. 2.

Таблиця 2

Модель класифікатора

Показник	SPAM	NOSPAM
Частота класів	2	1
Сума слів	6	3
Показую	1	0
Заробити	1	0
Допомога	1	0
Собаці	1	0
Як	1	1
Чапі	1	0
Привіт	0	1
Справи	0	1

Для перевірки класифікуємо текст «привіт, як справи». Розрахуємо значення виразу для класу SPAM:

$$\log\left(\frac{2}{3}\right)+\log\left(\frac{1}{8+6}\right)+\log\left(\frac{2}{8+6}\right)+\log\left(\frac{1}{8+6}\right)=-7{,}629.$$

Тепер виконаємо те ж саме для класу NOSPAM:

$$\log\left(\frac{1}{3}\right)+\log\left(\frac{2}{8+3}\right)+\log\left(\frac{2}{8+3}\right)+\log\left(\frac{1}{8+3}\right)=-6{,}906.$$

У цьому разі клас NOSPAM виграв, і листа з таким текстом не буде позначено як спам.

Висновки

1. У роботі проведено короткий огляд статистичних даних проблеми небажаної електронної кореспонденції. З огляду на статистичні дані чітко видно гостру проблему і великі масштаби затрат на опрацювання трафіку, пов'язаного з доставлянням спам-повідомлень.

2. Розглянуто три методи класифікації текстової інформації, їхні сильні і слабкі місця. У результаті аналізу методів найбільш привабливим для використання вибрано Naive Bayes Classifier, оскільки цей метод простий в імплементації та тестуванні, процес навчання досить ефективний порівняно з іншими більш складними класифікаторами і на невеликих обсягах документів різниця між іншими набагато складнішими алгоритмами класифікації часто неістотна, а іноді Naive Bayes Classifier може виявитися і більш точним. Також на прикладі розглянуто функціонування Naive Bayes Classifier.

Список використаної літератури

1. ***Guzella T. S., Caminhas W. M.*** *A review of machine learning approaches to Spam filtering // Expert Systems with Applications. 2009. Vol. 36, no. 7. P. 10206–10222. DOI:10.1016/j.eswa.2009.02.037.*

2. ***Metsis V., Androutsopoulos I., Paliouras G.*** *Spam Filtering with Naive Bayes — Which Naive Bayes? // CEAS 2006: Third Conference on Email and Anti-Spam (July 27-28, 2006). Mountain View, California, USA, 2006.*

3. ***A Bayesian*** *approach to filtering junk email / M. Sahami, S. Dumais, D. Heckerman, E. Horvitz // AAAI Workshop on Learning for Text Categorization. Technical Report. 1998.*

А. А. Пидмогильный, О. Н. Ткаченко, А. И. Голубенко, А. В. Дробик

NAIVE BAYES CLASSIFIER КАК ОДИН ИЗ ВАРИАНТОВ ФИЛЬТРАЦИИ НЕЖЕЛАТЕЛЬНОЙ ЭЛЕКТРОННОЙ ПОЧТЫ

Рассмотрена проблема классификации электронной корреспонденции для определения принадлежности писем к нежелательной электронной почты. Предложены три подхода для решения данной проблемы. Детально проанализирован Naive Bayes Classifier как наиболее эффективный способ, позволяющий с высокой точностью классифицировать письма и в то же время является экономическим решением относительно затрат вычислительных ресурсов.

Ключевые слова: Naive Bayes Classifier; вычислительная эффективность; Rule based classification; Weight based classification; фильтрация спам-писем; классификатор.

O. O. Pidmogylnyi, O. M. Tkachenkoaptev, O. I. Golubenko, O. V. Drobyk

NAIVE BAYES CLASSIFIER AS ONE WAY TO FILTER SPAM MAIL

In 2018-2019 years, mail spam was sent to the average for 53,8% of the previous traffic. Persons ranked highest in the ranking of spam is China (15,82%), USA (12,64%), Germany (5,86%), Russia (6,98%) and Brazil (6,95%). From these statistics we can calculate that on average electronic spending is about 0,17 terawatt, that is, about $282 billion is spent on spam e-mails. The article describes the problem of classifying e-mails to determine the affiliation of e-mails to spam. Three approaches are considered to solve this problem. This article considers three methods of text classification, their strengths, and weaknesses. As a result of the analysis of the methods, the most attractive to use is the Naive Bayes Classifier due to the fact that this method is easy to implement and test, the learning process is quite effective in comparison with other more complex classifiers and on small document cases the difference between other much more sophisticated classification algorithms is often irrelevant, and sometimes the Naive Bayes Classifier may be more accurate, as well as the example of how the Naive Bayes Classifier works, and it is considered in detail as the most efficient way to classify letters with high precision while being a cost-effective solution to detect spam messages.

Keywords: Naive Bayes Classifier; performance; Rule based classification; Weight based classification; spam filtering; classifier.