Контроль космічного та повітряного простору

УДК 621.384 **doi: 10.26906/SUNZ.2019.2.018**

Є. О. Гришманов[1], І. В. Захарченко[2], П. Г. Берднік[2], М. В. Кас'яненко[2]

[1] Льотна академія Національного авіаційного університету, Кропивницький, Україна
[2] Харківский національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна

ВИБІР МАТЕМАТИЧНОГО АПАРАТУ ДЛЯ ПОБУДОВИ ВЕКТОРНОЇ МОДЕЛІ ТЕКСТОВИХ ПОВІДОМЛЕНЬ ДЛЯ НАВЧАННЯ ГЛИБОКОЇ НЕЙРОННОЇ МЕРЕЖІ ПРОГНОЗУВАННЮ НЕСПРИЯТЛИВИХ АВІАЦІЙНИХ ПОДІЙ В ПОЛЬОТІ

В роботі проводиться дослідження і вибір математичного апарату для побудови словника і векторної моделі текстових повідомлень для навчання глибокої гібридної нейронної мережі прогнозуванню несприятливих авіаційних подій в польоті. Для визначення вагових значень слів в текстових повідомленнях про несприятливі авіаційні події в польоті при формуванні словника аналізуються вагові моделі на основі мір TF-IDF, TF-RF і TF-ICF. У якості методів векторного представлення текстової інформації в роботі досліджуються: «мішок слів», латентно-семантичний аналіз (Latent semantic analysis (LSA)), моделі векторного уявлення Word2Vec, Global Vectors (GloVe) та Doc2Vec. В результаті аналізу вказаних моделей і методів в якості базового підходу до формування словника уніграмм (біграмм) пропонується використовувати міру TF-ICF, а в якості моделі векторного уявлення слів (словосполучень) пропонується використовувати модель CBOW.

Ключові слова: безпека польотів, прогнозування, векторна модель текстових повідомлень, глибока нейронна мережа, CBOW, TF-ICF.

Вступ

Постановка проблеми. На сучасному етапі одним з необхідних елементів забезпечення безпеки польотів є застосування автоматизованих систем прогнозування несприятливих авіаційних подій під час польоту. Слід зазначити, що при розробці алгоритмів та методів прогнозування авіаційних подій необхідно враховувати той факт, що дана прикладна область постійно зазнає змін, отже ускладнюється об'єкт дослідження, що в свою чергу вимагає застосування сучасних інформаційних технологій.

Перспективним підходом для вирішення задач прогнозування є використання глибоких гібридних нейронних мереж, які на відміну від інших методів дозволяють враховувати велику кількість факторів, поданих не тільки у кількісному, а й у якісному вигляді, дозволяють здійснювати моделювання при невеликих експериментальних вибірках та мають здатність до навчання [1–3].

Для навчання глибокої гібридної нейронної мережі прогнозуванню несприятливих авіаційних подій в польоті необхідно сформувати навчальну вибірку, а саме: побудувати таку модель текстових повідомлень, яка б описувала певний клас авіаційної події і могла розглядатися у вигляді структури даних для подачі на вхід глибокої гібридної нейронної мережі.

В якості такої структури даних прийнято розглядати векторні представлення текстових повідомлень та слів (словосполучень) [4].

Мета статті. Проведення дослідження та вибір ефективного математичного апарату для побудови векторної моделі текстових повідомлень для навчання глибокої нейронної мережі прогнозуванню несприятливих авіаційних подій в польоті.

Основний матеріал

При побудові моделі текстових повідомлень в роботі прийняті такі обмеження та припущення:

в якості текстових повідомлень в роботі розглядається як неструктурований, так і структурований набір даних у вигляді коротких речень (загальний розмір текстового повідомлення до 1000 слів) з великою кількістю навчальних прикладів;

перед безпосередньою побудовою векторної моделі текстових повідомлень створюються два словника: словник використання окремих слів (уніграмм) і словник використання комбінації (словосполучення) двох слів (біграмм);

формування навчальної вибірки для навчання глибокої гібридної нейронної мережі розглядається в рамках виконання задачі N-арної класифікації (за кількістю класів несприятливих авіаційних подій в польоті).

Модель навчальної вибірки M_{acc} представимо у вигляді векторного представлення слів (словосполучень), яка має наступний вигляд:

$$M_{acc} : \{K_s\} \to \{D_l\} \quad (1)$$

де D_l - формалізоване l-те текстове повідомлення про авіаційну подію у вигляді вектора $(w_1, w_2, ...w_V)$; w_i – вага елементу текстового повідомлення, що розглядається (слово (уніграмма), словосполучення (біграмма)); V – множина унікальних елементів тестового повідомлення (уніграмм, біграмм).

Векторне представлення слів (словосполучень) дозволяє значно покращити якість методів автоматичної обробки текстових повідомлень з використанням нейронних мереж.

В якості текстових повідомлень в роботі розглядається як неструктурований, так і структурований набір даних у вигляді коротких речень (загальний розмір текстового повідомлення до 1000 слів) з великою кількістю навчальних прикладів. Перед побудовою векторної моделі текстових повідомлень про несприятливі авіаційних події в польоті створюються два словника: словник використання окремих слів (уніграмм) і словник використання комбінації (словосполучення) двох слів (біграмм).

При аналізі текстових повідомлень з використанням глибоких нейронних мереж є необхідність подання слів (словосполучень) текстового повідомлення в вигляді векторів. В даному випадку в якості однослівних уніграмм виділяються всі слова текстового повідомлення, у якості біграмм - унікальні комбінації двох слів для додаткового опису особливостей авіаційної події. За результатами попередньої обробки текстових повідомлень в подальшому не розглядаються прийменники, розділові знаки, власні імена. В даний час існують різні підходи до вилучення ключових оціночних слів (словосполучень) з текстів та визначення їх ваги в наборі текстових повідомлень [4].

В якості базового підходу до вилучення ключових оціночних слів (словосполучень) для формування словника текстових повідомлень про несприятливі авіаційних події в польоті пропонується використовувати N-грамові або вагові моделі векторного кодування [4]. Дані моделі дозволяють забезпечити автоматичне або напівавтоматичне вилучення слів (словосполучень) з текстів, зменшити вагу широко уживаних слів і збільшити вагу більш рідкісних слів, які можуть досить точно вказати на те, до якого класу належить текст, заснований на значущості цього слова для набору текстових повідомлень.

Використання N-грам в загальному випадку зводиться до побудови для кожного класу вектору одиничної норми, що відображає частоту виникнення різних N -грамм в текстових повідомленнях.

Побудова відповідного вектора на основі N-грамм включає виконання наступних етапів:

1. Для кожного класу формується множина, що складається з N-грам (слів або словосполучень), які найбільш часто зустрічаються в текстових повідомленнях про даний клас.

2. Формується множина, що складається з найбільш частих N-грам з об'єднанням всіх множин, отриманих на першому етапі.

3. Для кожного класу формується відповідний вектор – М-мірний вектор одиничної норми, що відображає частоту зустрічання N -грамм з множини, отриманої на другому етапі.

У загальному випадку результуючі вектори подаються на вхід класифікатора, наприклад SVM (Support Vector Machine) або XGBoos [5]. Однак дане векторне подання не підходить для нейромережевих класифікаторів, до яких пред'являються вимоги по обмеженню розміру словника. В даному дослідженні моделі N-грамм розглядаються тільки в контексті побудови словника, а саме за результатами визначення вагових значень слів в текстових повідомленнях про несприятливі авіаційнї події в польоті зі словника видаляються низькочастотні слова, тобто слова, що дуже рідко зустрічаються в текстових повідомленнях.

Даний підхід дозволяє зменшити обчислювальну складність як при побудові векторної моделі текстових повідомлень, так безпосередньо при використанні нейромережевого класифікатора несприятливих авіаційних подій в польоті.

Розглянемо три основних вагових моделі на основі мір TF-IDF, TF-RF і TF-ICF [5]. Найбільш поширена модель на основі міри TF-IDF (Term Frequency - Inverse Document Frequency), яка визначається відповідно до наступного виразу [5]:
TF – частота зустрічі слова (словосполучення) в текстовому повідомленні по класах несприятливих авіаційних подій в польоті; *T* – – загальна кількість текстових повідомлень по класах несприятливих авіаційних подій в польоті; $T(t_i)$ – кількість повідомлень, які містять слово(словосполучення), що розглядається. Основна ідея використання міри *TF-RF* (Term Frequency - Relevance Frequency) полягає в тому, що вага слова (словосполучення) обчислюється на основі інформації про розподіл цього слова в текстових повідомленнях і при цьому враховується приналежність текстових повідомлень до двох заданих класів. Міра *TF-RF* розраховується у відповідності до наступного виразу:

$$V_{TF-RF} : TF \times \log\left(2 + a/\max(1,c)\right), \qquad (2)$$

де *a* – кількість текстових повідомлень 1-го класу подій, що містить зважуване слово; *c* – кількість текстових повідомлень 2-го класу подій, що містить зважуване слово.

Міра *TF-ICF* (Term Frequency – Inverse Category Frequency) визначається згідно виразу[5]:

$$V_{TF-ICF} : TF \times \log\left(1 + |C|/cf(ti)\right) \qquad (3)$$

де *c* – кількість класів в даній предметній області; *cf* – кількість класів, де зустрічається зважуване слово.

Основна ідея міри *TF- ICF* полягає в тому, що вага слова обчислюється на основі інформації про розподіл цього слова в текстових повідомленнях і враховує приналежність текстових повідомлень до певних класів.

Міра *TF-IDF* (Term Frequency Inverse Document Frequency) — статистична міра, що використовується для оцінювання важливості слова в контексті повідомлення, що є частиною колекції повідомлень. Основна ідея міри *TF-IDF* полягає в тому, що вага деякого слова пропорційна частоті вживання цього слова у текстовому повідомленні та зворотньо пропорційна частоті вживання даного слова у всіх повідомленнях класів.

$$V_{TF-IDF} : TF \times \log\left(1 + |C| \Big/ \left|\{c_i \in C | t \in c_i\}\right|\right), \qquad (4)$$

де *c* – кількість класів в даній предметній області;

$\{c_{i\in} C | t \in c_i\}$ – кількість класів з колекції *C*, де зустрічається слово (словосполучення)t.

Надалі при формуванні словників для використання окремих слів (уніграмм) і комбінації двох слів (біграмм) в роботі пропонується використовувати модифіковану міру *TF- ICF*. Це пов'язано з тим, що моделі вилучення слів (словосполучень), засновані на мірі *TF- ICF*, при багатовимірній класифікації показують результати краще, ніж моделі, засновані на мірах *TF-IDF* та *TF-RF* [6].

У якості методів векторного представлення текстової інформації в роботі досліджуються «мішок слів» (Bag of words) [4], латентно-семантичний аналіз (Latent semantic analysis (LSA)) [4], моделі векторного уявлення Word2Vec [5], Global Vectors (GloVe) [5] і Doc2Vec [6].

«Мішок слів» - модель, в якій текстові повідомлення мають вигляд неупорядкованого набору слів без зв'язку між ними. Дана модель навчається на словнику, складеному зі слів будь-яких текстових повідомлень. Серед основних недоліків даної моделі можна виділити такі: дуже великий розмір векторів; уповільнення операції порівняння векторів через їх розмірність; можливість застосування різних методів зниження розмірності призводить до втрати якості.

Латентно-семантичний аналіз (LSA) представляє собою ефективний статистичний алгоритм, що складається з двох основних етапів: побудова термів документної матриці і виконання сингулярного розкладення. Основою методу латентно-семантичного аналізу є принципи факторного аналізу, зокрема виявлення латентних зв'язків об'єктів, що вивчаються.

Даний метод ґрунтується на декількох параметрах, таких як: локальні і глобальні частоти зустрічі слів, функції локального та глобального зважування і розмірність семантичного простору.

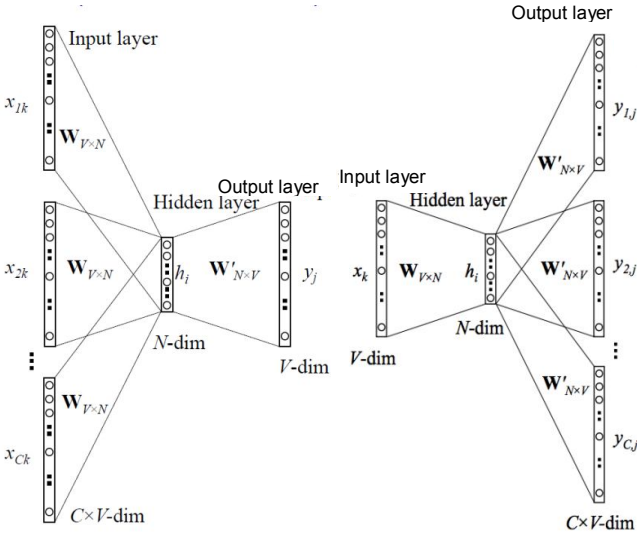
В рамках даного дослідження розглянемо технологію Word2Vec [7], що заснована на дистрибутивній семантиці та векторному представленні слів. Для Word2Vec відомі два алгоритми навчання: безперервний мішок слів (Continuous Bag of Words (CBOW)); Skip-gram.

Модель CBOW - модель мішка слів, що враховує чотири найближчих сусіда (два попередніх і два наступні слова), при цьому порядок проходження слів не враховується. Принципом роботи CBOW є передбачення слова при заданому контексті. У CBOW використовуються три шари. Вхідний шар відповідає контексту. Прихований шар - проекції кожного слова з вхідного шару в вагову матрицю, яка проектується в третій вихідний рівень. Останнім етапом моделі є порівняння її виведення з самим словом, щоб скорегувати його представлення, засноване на зворотному поширенні градієнта помилки. Узагальнена структурна схема моделі CBOW представлена на рис. 1, а. Метою виконання моделі CBOW є максимізація виразу:

$$\sum_{t=1}^{V} \log p\left(m_t \middle| m_{t-c/2} \ldots m_{t+c/2}\right) \Big/ V, , \qquad (5)$$

де V – розмір словника; m_t – t-й елемент словника; c – розмір вікна для кожного слова (словосполучення).

У моделі Skip-Gram вирішується зворотна задача − на підставі одного слова передбачається контекст. Останній крок алгоритму – порівняння виведення з кожним словом в контексті з метою коригування уявлення, заснованого на зворотному поширенні градієнта помилки.



а б

Рис. 1. Узагальнена структурна схема моделі:
а – CBOW; б – Skip-gram

Архітектура типу Skip-gram має інший принцип: вона використовую поточне слово, для того щоб передбачати інші слова, що його оточують.

Основна ідея Skip-Gram полягає в максимізації класифікації слова, ґрунтуючись на іншому слові в цьому ж реченні. Даний метод виконує максимізацію наступного виразу:

$$\sum_{t=1}^{V} \sum_{j=t-c, j\neq t}^{t=c} \log p(m_j \middle| m_t) \Big/ V, \qquad (6)$$

Узагальнена структурна схема моделі Skip-gram наведена на рис. 1, б.

Порядок слів не впливає на результат в жодному з цих алгоритмів.

Результати досліджень, проведені в роботі [7], свідчать, що модель CBOW ефективна для невеликих наборів даних та краще обробляє слова (короткі речення, але велика кількість прикладів), що часто зустрічаються. При цьому модель CBOW характеризується високою швидкодією, що є особливо важливим при прогнозуванні авіаційних подій в реальному часі. Модель Skip-gram в свою чергу більш ефективна на великих наборах даних, за допомогою її добре описуються слова (довгі речення, але прикладів набагато менше), що рідко зустрічаються. В основі методу GloVe лежить спосіб підрахунку частоти появи слів в безлічі тестових повідомлень [8]. Реалізація методу виконується в два етапи: на першому етапі виконується побудова матриці суміжності з навчальної множини, на другому етапі−факторизація матриці суміжності для отримання векторів.

Модель Doc2Vec заснована на алгоритмі навчання без учителя. Вчиться отримувати розподілені вектори для частин текстів. У даній моделі векторні уявлення текстових повідомлень навчаються передбачати наступне слово з урахуванням контексту.

Вектори слів і текстових повідомлень навчаються з використанням методу стохастичного градієнтного спуску і методу зворотного поширення помилки. Вектори текстових повідомлень є унікальними, а вектори однакових слів у різних текстових повідомленнях збігаються.

Згідно з обмеженнями і припущеннями, наведеними вище, в якості базового підходу до формування словника уніграмм (биграмм) пропонується використовувати міру TF-ICF (з урахуванням рішення задачі багатовимірної класифікації), а в якості моделі векторного уявлення слів (словосполучень) пропонується використовувати модель CBOW (з урахуванням вимог до повноти й оперативності обробки текстових повідомлень про несприятливі авіаційних події в польоті).

Висновки

В роботі проведено аналіз методів, за допомогою яких можливе вирішення задачі побудови словника і векторної моделі текстових повідомлень для навчання глибокої гібридної нейронної мережі прогнозуванню несприятливих авіаційних подій в польоті. В результаті проведеного аналізу було встановлено, що для побудови словника уніграмм (біграмм) в якості базового математичного апарату доцільним є використання міри TF-ICF (з урахуванням вирішення задач багатомірної класифікації), а в якості моделі векторного представлення слів (словос-сполучень) доцільним є використання моделі CBOW (з урахуванням вимог до повноти та оперативності обробки текстових повідомлень про несприятливі авіаційні події в польоті).

СПИСОК ЛІТЕРАТУРИ

1. Григорків В.С. Нейронні мережі та їхнє використання для прогнозування тенденцій ринку нерухомості // В.С. Григорків, О.І. Ярошенко, Н.В. Філіпчук / Науковий вісник НЛТУ України. – 2012. – Вип. 22.5. – С. 328-33.
2. Y. Kim. Convolutional neural networks for sentence classification. arXiv:1408.5882 [cs.CL], 2014.
3. C. Olah. Neural networks, recurrent neural networks, convolutional neural networks. Ел. ресурс/ http://colah.github.io.htm/
4. Крейнес М. Г. Модели текстов и текстовых коллекций для поиска и анализа информаци // М. Г. Крейнес / Матем. модел. эколого-экономич. систем: экономика ТРУДЫ МФТИ. – 2017. – Том 9(3). – С. 132-142.
5. Reed J.W., Jiao Y., Potok T.E., Klump B.A., Elmore M.T., Hurson A.R. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams // In: Proc.Machine Learning and Applications (ICMLA '06). 2006. pp. 258–263.
6. П.Флах Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из даннях / пер. с англ А.А.Слинкина. − М.: ДМК Пресс, 2015.− 400 с.
7. Mikolov T. Distributed representations of words and phrases and their compositionality / T.Mikolov, I.Sutskever, K.Chen, G.S. Corrado, J. Dean // Advances in neural information processing systems. 2013. P. 3111–3119.
8. Борисов Е.С.Автоматизированная обработка текстов на естественном языке, с использованием инструментов языка Python /Електронний ресурс/ http://mechanoid.kiev.ua/ml-text-proc.htm.

Рецензент: д-р техн. наук, проф. О. І. Тимочко
Харківський національний університет Повітряних Сил імені Івана Кожедуба
Received (Надійшла) 15.01.2019
Accepted for publication (Прийнята до друку) 20.03.2018

Выбор математического аппарата для построения векторной модели текстовых сообщений для обучения глубокой нейронной сети прогнозированию неблагоприятных авиационных событий в полете

Е. А. Гришманов, И. В. Захарченко, П. Г. Бердник, М. В. Касьяненко

В работе проводится исследование и выбор математического аппарата для построения словаря и векторной модели текстовых сообщений для обучения глубокой гибридной нейронной сети прогнозированию неблагоприятных авиационных событий в полете. Для определения весовых значений слов в текстовых сообщениях о не- благоприятных авиационных происшествиях в полете при формировании словаря анализируются весовые модели на основе мер TF-IDF, TF-RF и TF-ICF. В качестве методов векторного представления текстовой информации в работе исследуются: «мешок слов», латентно-семантический анализ (Latent semantic analysis (LSA)), модели векторного представления Word2Vec, Global Vectors (GloVe) и Doc2Vec. В результате анализа указанных моделей и методов в качестве базового подхода к формированию словаря униграмм (биграммы) предлагается использовать меру TF-ICF, а в качестве модели векторного представления слов (словосочетаний) предлагается использовать модель CBOW.

Ключевые слова: безопасность полетов, прогнозирование, векторная модель текстовых сообщений, гибридная нейронная сеть, Word2Vec, CBOW, TF-ICF.

Choice of a mathematical instrument for constructing a vector text message model for training a deep neural network to predict unfavorable aircraft accidents in the flight

E. Grishmanov, I. Zakharchenko, P. Berdnik, M. Kasyanenko

The paper studies and selects a mathematical instrument for constructing a dictionary and a vector model of text messages for teaching a deep hybrid neural network to predict unfavorable aircraft accidents in the flight. To determine the weight values of words in text messages about unfavorable aircraft accidents in the flight during the formation of the dictionary, weighting models based on the measures TF-IDF, TF-RF and TF-ICF are analyzed. As methods of vector representation of text information, the paper analyzes: “bag of words”, latent-semantic analysis and models of vector representation, such as Word2Vec, Global Vectors (GloVe) and Doc2Vec. As a result of the analysis of these models and methods, it is proposed to use the TF-ICF measure as the basic approach to the formation of the unigram vocabulary (bigrams), and use the CBOW model as a model for the vector representation of words (word combinations).

Keywords: flight safety, prediction, text messaging vector model, hybrid neural network, Word2Vec, CBOW, TF-ICF.