Вісник Національного технічного університету "ХПІ", 2018, № 42 (1318)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.852 **DOI: 10.20998/2411-0558.2018.42.17**

А. Ю. ТІТОВА, канд. техн. наук, Державний університет телекомунікацій, Київ,
Д. Є. ІВАНОВ, д-р техн. наук, доц., ІПММ НАНУ, Слов'янськ

РОЗРОБКА МОДЕЛІ АНАЛІЗУ СКЛАДНИХ ДАНИХ НА ОСНОВІ КЛАСИФІКАЦІЇ MACHINE LEARNING

Виконано аналіз методів класифікації machine learning та визначені етапи обробки складних даних на основі бінарної класифікації. Розроблено модель аналізу складних даних на основі класифікації machine learning та проведено перевірку її адекватності з використанням різних засобів оцінки. Виконана класифікація даних на відповідність двом класам: корисної інформації та спаму. Іл.: 2. Бібліогр.: 11 назв.

Ключові слова: класифікація; складні дані; machine learning; засоби оцінки; бінарна класифікація; спам.

Постановка проблеми. Проблема аналізу даних складної структури є актуальною у сучасному світі інформаційних технологій для класифікації даних, прогнозування кількісних характеристик, виявлення нових закономірностей та інтерпретації даних. Сучасні методи класифікації та технології машинного навчання, які використовуються для аналізу та прогнозування [1, 2], дають високі показники точності, швидкодії, проте виявлення цінних даних, програмна реалізація та експериментальне впровадження результатів недостатньо висвітлені авторами. Для вирішення проблеми обробки даних складної структури слід розробити модель аналізу складних даних на основі класифікації machine learning та застосовувати прогресивні мови програмування для експериментального дослідження

Аналіз літератури. Сучасні програмні платформи та середовища розробки дозволяють реалізовувати моделі та методи різної складності для класифікації, прогнозування даних та розробки інформаційних технологій. Для класифікації зображень місцевості обрані чотири алгоритми класифікації, а саме: дерева рішень, наївний метод Байєса (NB), випадкові ліси та машини опорних векторів (SVM), останній показав потенціал у платформі Haddop MapReduce та свою продуктивність [1]. Розроблено модель на основі програмного агента, що дозволяє класифікувати пацієнтів із захворюванням на цукровий діабет на групи та прогнозувати необхідну кількість медичних препаратів для конкретного випадку [2]. У дослідженнях [3 – 5] розглядаються методи NB та TAN, опорних векторів (SVM), *k*-найближчих сусідів, дерев

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

рішень для прогнозування кредитоспроможності фізичних осіб, для керування складними електромеханічними системами, нелінійними об’єктами, а також об’єктами зі стохастичними параметрами, в задачах класифікації текстових документів та обрані методи з достатніми показниками точності класифікації для відповідних задач. Проведено дослідження особливостей класифікації методів і технологій аналітики Великих даних та технологій business intelligence під час обробки діагностичної інформації для збереження конкурентоспроможності підприємств [6, 7].

Визначено основні переваги методів глибинного навчання над традиційними підходами до задач класифікації, для відокремлення ознак із супутникових даних [8]. Запропоновані підходи до моделювання кластеризації та класифікації функціональних, багатомірних даних, що містять неоднорідність, відсутність інформації та динамічну приховану структуру з декількох областей застосування [9].

Існуючі методи та моделі класифікації складних даних використовують математичний апарат для конкретної задачі та дають не значні показники точності, тому в даній роботі запропоновано розробити нову модель аналізу складних даних.

Мета дослідження – розробка моделі аналізу складних даних на основі класифікації machine learning для прогнозування кількісних характеристик, дійсних і придатних до подальшого використання даних.

Для досягнення даної мети необхідно вирішити наступні задачі:

– визначити етапи обробки складних даних на основі бінарної класифікації;

– виконати оцінку адекватності моделі аналізу складних даних різними засобами на конкретних прикладах.

Матеріали дослідження. Відомо, що класифікація є популярною задачею машинного навчання, та полягає у побудові моделей, що виконують віднесення обраного об'єкта до одного з декількох відомих класів [10]. Одним з головних недоліків методу дерев рішень для задач класифікації текстів є те, що "позитивні" і "негативні" розгалуження у вузлах мають однакову вагу. До переваг зазначеного методу слід віднести той факт, що побудоване дерево легко піддається аналізу. Результати класифікації текстів за допомогою методу опорних векторів є одними з найкращих, у порівнянні з іншими методами машинного навчання. Однак, швидкість навчання SVM одна з найнижчих. Для проведення дослідження обрано бінарну класифікацію.

Об’єктом даного дослідження є таблиця даних із електронних листів, що містять спам, котрий розміщений у різних комірках. Такі дані

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

розміщено у колекції наборів даних Центру машинного навчання та інтелектуальних систем Каліфорнійського університету (Center for Machine Learning and Intelligent Systems). Необхідно виконати класифікацію даних на відповідність двом класам, а саме корисній інформації та спаму.

Для цього слід виконати наступні етапи:

1. Завантажити файл даних для аналізу у середовище.

2. Розділити вихідні дані в співвідношенні 10:1 на навчальну і перевірочну вибірки.

3. Використати логістичну регресійну модель для класифікації даних.

4. Виконати різними засобами оцінку якості моделі аналізу складних даних на конкретних прикладах.

Для бінарної класифікації табличних даних запропоновано використати логістичну регресійну модель, що дозволяє на основі отриманих остач від прогнозу, визначити кількість корисних даних та спаму для навчальної та перевірочної вибірок.

Для оцінки адекватності моделі аналізу складних даних використано різні засоби, а саме кількісні показники прогнозу; ROC-криву для оцінки ймовірності спаму; графік щільності розподілу ймовірностей обох класів (спаму та корисної інформації).

Для розрахунку кількісних показників класифікації табличних даних на корисні та зі спамом отримано чутливість, специфічність та точність прогнозу.

Розрахувати чутливість (sensitivity), що визначає наскільки вдало виявлено дані зі спамом, необхідно за наступним виразом [11]

$$SE = TP / (TP + FN), \quad (1)$$

де TP – число даних з істинно позитивним результатом; FN – кількість даних з хибно негативним результатом прогнозу.

Для представлення ефективності класифікації, а саме відповідності правильного виявлення спаму від корисної інформації, необхідно обчислити специфічність (specificity) за наступним виразом:

$$SP = FP / (FP + TN), \quad (2)$$

де FP – число даних з хибно позитивним результатом прогнозу; TN – число даних з істинно негативним результатом.

Необхідно визначити загальну ймовірність прогнозу давати правильні результати, для цього розрахувати точність (accuracy) за виразом [11], що представлено далі:

$$AC = (TP + TN) / (TP + FP + FN + TN), \quad (3)$$

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Після проведення експерименту отримано такі значення для кількісних показників прогнозу:

$$SE = 0.89;\ SP = 0.86;\ AC = 0.93.$$

Для оцінки якості прогнозу графоаналітичним методом та інтерпретації перерахованих показників необхідно застосувати ROC-аналіз. Побудовано ROC-криву (рис. 1) для бінарного відгуку (1 – спам, 0 – корисна інформація), де довільне значення даних на цьому діапазоні вважається класифікаційним порогом. Чим ближче крива до верхнього лівого кута, тим вище у моделі здатність до прогнозу.

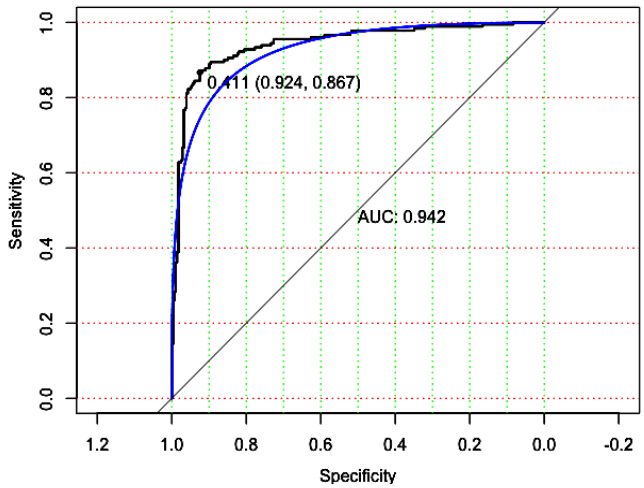

Рис. 1. ROC-крива оцінки класифікації моделі

Запропонована логістична регресійна модель дозволяє отримати прогноз класу кожного із набору складних даних та повернути оцінену ймовірність належності даних відповідному класу. При підборі оптимальних порогових значень класифікатора моделі проаналізовано графік щільності розподілу ймовірностей обох класів, котрий представлений на рис. 2.

Після проведення експерименту отримано такі результати показників прогнозу:

$$SE = 0.8864629;\ SP = 0.8645161;\ AC = 0.9324324.$$

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

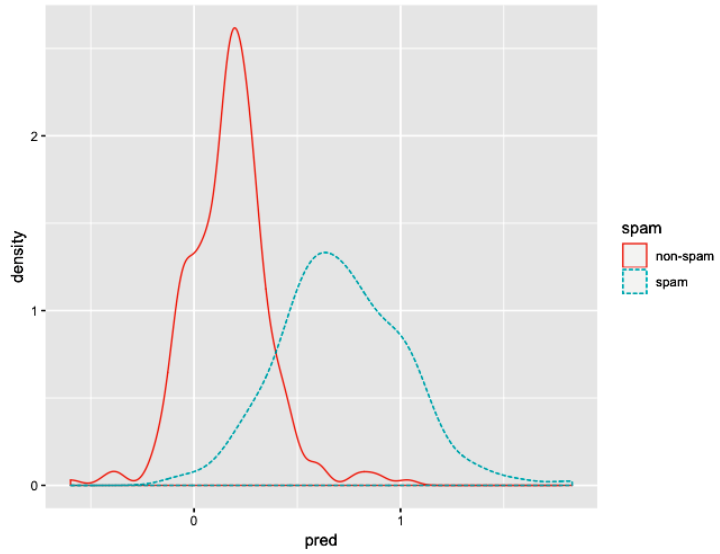

Рис. 2. Графік щільності розподілу ймовірностей появи даних двох класів

Як видно з рис. 2, кривими лініями показана умовна ймовірність прогнозу корисної інформації та спаму у тестовій вибірці складних даних, де на осі *X* відображено значення ймовірностей прогнозу, а на осі *Y* – щільність розподілу даних між двома класами.

Висновки. Під час дослідження проаналізовані методи класифікації machine learning, визначені етапи обробки складних даних на основі бінарної класифікації; розроблено модель аналізу складних даних, виконано експериментальні дослідження застосування моделі на конкретних прикладах. Отримані результати свідчать про можливість використання моделі.

Список літератури:

1. *Ayma V.A.* Classification algorithms for big data analysis, a Map Reduce approach / *V.A. Ayma, R.S. Ferreira, P. Happ, D. Oliveira* // The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2015. – Т. 40. – № 3. – P. 17-21.

2. *Alotaibi N.M.* Agent-based big data classification / *N.M. Alotaibi. M. Abdullah. H. Mosli* // Journal of Fundamental and Applied Sciences. – 2018. – Vol. 10. – №. 4. – P. 258-264.

3. *Кириченко В.Е.* Застосування наївного та дереводоповненого байєсівських класифікаторів для прогнозування кредитоспроможності фізичних осіб / *В.Е. Кириченко, О.М. Терентьєв, Н.О. Связінська* // System analysis and information technology: Proceedings of 18-th International conference SAIT 2016, Kyiv, Ukraine, May 30 – June 2, 2016. – NTUU "KPI", 2016. – С. 364-365.

4. *Шеремет О.І.* Метод опорних векторів (SVM) / *О.І. Шеремет, О.В. Садовой* // Математичне моделювання. Науковий журнал. Дніпродзержинськ: ДДТУ. – 2013 – № 1 (28). – С. 13-17.

5. *Волосюк Ю.В.* Методи класифікації текстових документів в задачах Text Mining / *Ю.В. Волосюк* // Наукові записки Українського науково-дослідного інституту зв'язку. – 2014. – №. 6. – С. 76-81.
6. *Верес О.М.* Класифікація методів аналізу Великих даних / *О.М. Верес, Р.М. Оливко* // Вісник Національного університету "Львівська політехніка". Серія: Інформаційні системи та мережі. – Львів :Видавництво Львівської політехніки, 2017. – № 872. – С. 84–92.
7. *Титова А.Ю.* Анализ технологий business intelligence при обработке диагностической информации / *А.Ю.* Титова // Материалы Регионального семинара Международного союза электросвязи для стран Европы и СНГ "Цифровое будущее на основе 4G/5*G*", г. Киев, 14-16 мая 2018. – 2018. – С. 84-85
8. *Лавренюк М.С.* Огляд методів машинного навчання для класифікації великих обсягів супутникових даних / *М.С. Лавренюк, О.М. Новіков* // Системні дослідження та інформаційні технології. – 2018. – №. 1. – С. 52-71.
9. *Chamroukhi F.* Model-Based Clustering and Classification of Functional Data / *F. Chamroukhi, H.D. Nguyen* // Cornell University Library: Statistics-Machine Learning. – 2018. – 69 P: available at: https://arxiv.org/abs/1803.00276v2 (accessed December 2018).
10. *Шитиков В.К., Мастицкий С. Э.* Классификация, регрессия, алгоритмы Data Mining с использованием R [Электронный ресурс] – Режим доступа: https://ranalytics.github.io/data-mining/index.html (accessed December 2018).
11. *Тітова А.Ю.* Методи та моделі інформаційної технології для автоматизованих систем переробки діагностичної інформації на основі термограм: автореф. дис. ... канд. техн. наук. – Київ. – 2017. – 23 с.

References:

1. Ayma, V.A., Ferreira, R.S., Happ, P., and Oliveira, D. (2015), "Classification algorithms for big data analysis, a Map Reduce approach", *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 40, No. 3, pp 17-21.
2. Alotaibi, N.M., Abdullah, M., and Mosli, H. (2018), "Agent-based big data classification", *Journal of Fundamental and Applied Sciences*, Vol. 10, No. 4., pp. 258-264.
3. Kirichenko, V.E., Terentyev, O.M., and Svyazninskaya, N.O. (2016), "Application of naive and wood-based Bayesian classifiers for predicting the creditworthiness of individuals", *System analysis and information technology: Proceedings of 18-th International conference SAIT 2016*, NTUU "KPI", Kyiv, Ukraine, May 30 – June 2, 2016, pp. 364-365.
4. Sheremet, A.I, and Garden, A.V. (2013), "Support vector machine (SVM)", *Mathematical Modeling. Scientific. Journal. Dneprodzerzhinsk: DonSTU*, No. 1 (28), pp. 13-17.
5. Volosyuk Yu.V. (2014), "Methods of classification of text documents in tasks Text Mining", *Scientific notes Ukrainian Research Institute of Communications*, No. 6, pp. 76-81.
6. Veres. O.M., and Olivko, R.M. (2017), "Classification of methods for the big data analytics", *Proceedings of the National University "Lviv Polytechnic". Series: Information systems and networks*, Vydavnytstvo Lvivskoi politekhniky, Lviv, No 872, pp. 84-92.
7. Titova A.Yu. (2018), "Analysis of business intelligence technologies in diagnostic information processing", *Proceedings of Regional Workshop of the International Telecommunication Union for Europe and CIS region "Digital Future Powered by 4G/5G"*, Kiev, May, 14-16, 2018, pp. 84-85
8. Lavreniuk, M., and Novikov, A. (2018), "Overview of machine learning to classify large volumes of satellite data", *System Research & Information Technologies*, No. 1, pp. 52-71.
9. Chamroukhi, F., and Nguyen, H.D. (2018), "Model-Based Clustering and Classification of Functional Data", *Cornell University Library: Statistics-Machine Learning*, 69 P, available at: https://arxiv.org/abs/1803.00276v2 (accessed December 2018).

10. Shitikov, V.K., and Mastitsky, S.E. (2017), "Classification, regression, Data Mining algorithms using R", available at: https://ranalytics.github.io/data-mining/index.html (accessed December 2018).
11. Titova, A.Yu. (2017), *Methods and models of information technology for an automated system for processing diagnostic information on the basis of thermal images: Author's thesis.* Kiev, 23 p.

Статтю представив д-р техн. наук, проф. ДУТ Вишнівський В.В.

Надійшла (received) 06.11.2018

Titova Anastasiya, PhD Tech
State University of Telecommunications
Str. Solomenska, 7, Kyiv, Ukraine, 03110
Tel.: (095) 333-51-01, e-mail: a.titova.wk@gmail.com
ORCID ID: 0000-0002-4803-2090

Ivanov Dmitry, Dr.Sci.Tech, Ass. Professor
Institute of Applied Mathematics and Mechanics
Str. Gen. Batyuka, 19, Slavyansk, 84100
Tel: (063) 559-51-90, e-mail: dmitry.ivanov.iamm@gmail.com
ORCID ID: 0000-0001-9956-6589

УДК 004.852

Розробка моделі аналізу складних даних на основі класифікації machine learning / Тітова А.Ю., Іванов Д.Є. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 42 (1318). – С. 171 – 178.

Виконано аналіз методів класифікації machine learning та визначені етапи обробки складних даних на основі бінарної класифікації. Розроблено модель аналізу складних даних на основі класифікації machine learning та проведено перевірку її адекватності з використанням різних засобів оцінки. Виконана класифікація даних на відповідність двом класам: корисної інформації та спаму. Іл.: 2, Бібліогр.: 11 назв.

Ключові слова: класифікація; складні дані; machine learning; засоби оцінки; бінарна класифікація; спам.

УДК 004.852

Разработка модели анализа сложных данных на основе классификации machine learning / Титова А.Ю., Иванов Д.Е. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2018. – № 42 (1318). – С. 171 – 178.

Выполнен анализ методов классификации machine learning и определены этапы обработки сложных данных на основе бинарной классификации. Разработана модель анализа сложных данных на основе классификации machine learning и проведена проверка ее адекватности с использованием различных средств оценки. Выполнена классификация даних на соответствие двум классам: полезной информации и спама. Ил.: 2, Библиогр.: 11 назв.

Ключевые слова: классификация; сложные данные; machine learning; средства оценки; бинарная классификация; спам/

УДК 004.852

Development of model analysis of a complex data based on machine learning classification / Titova A.Yu., Ivanov D.E. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2018. – № 42 (1318). – P. 171 – 178.

The analysis of methods of classification of machine learning and defined stages of processing of complex data on the basis of binary classification is performed. A model for analyzing complex data based on machine learning has been developed and a validation of its adequacy has been carried out using various means of evaluation. A classification of dyne has been performed to correspond to two classes: useful information and spam. Fig.: 2, Refs.: 11 titles.

Keywords: classification; complex data; machine learning; evaluation tools; binary classification; spam.