РЕФЕРАТ

Актуальність теми Щодня об’єми даних в мережі зростатють і працювати з ними стає дедалі складніше. На допомогу ручній обробці та аналізу даних приходить автоматизація та класифікація за допомогою алгоритмів машинного навчання. Аналогічно зростає кількість користувачів, що хоче мати доступ до цих даних у зручному форматі, поділеному на категорії та у впорядкованому вигляді. Саме для цього і були розроблені алгоритми класифікації та кластеризації, що допомагають комп’ютеру здійснити дані процеси. З точки зору науковців, що займаються обробкою даних та розробкою алгоритмів класифікації, дані алгоритми є не завжди оптимальними і вимагають додаткових тонких налаштувань для отримання кращих результатів. Отже, цілком логічним є існування попиту на ринку для розробки універсальних підходів, що допоможуть здійснювати класифікацію для різних вхідних даних, широкого кола алгоритмів та найголовніше будуть зрозумілими не тільки для розробника алгоритму, але й для пересічного користувача.

Об’єктом дослідження є процес побудови алгоритму універсальної прогностичної моделі.

Предметом дослідження є методи побудови прогностичних моделей та алгоритми класифікації даних.

Мета роботи: створення нового алгоритму побудови прогностичної моделі, що буде демонструвати точність передбачення не меншу, ніж аналогічні моделі для схожого роду вхідних даних, та мати просту реалізацію.

Методи дослідження. В роботі використовуються методи збору даних, методи класифікації текстових даних та статистичні методи.

Наукова новизна роботи полягає в наступному:

1. Запропоновано підхід, результатом якого є універсальна прогностична модель, що дає змогу абстрагуватися від конкретних реалізацій і використовувати її для тих самих даних з аналогічними показниками точності та кращими показниками швидкодії.

2. Наведено процес перетворення будь-якої прогностичної моделі чи деякої композиції моделей для перетворення в універсальну модель.

3. Підтверджено значно більші показники швидкодії моделі, розробленої за допомогою даного підходу.

На даному етапі роботи отриманих даних досить для того, щоб почати використовувати даний підхід для роботи з реальними даними та заміною існуючих алгоритмів.

Практична цінність отриманих в роботі результатів полягає в тому, що запропонований метод побудови прогностичної моделі дозволяє збільшити швидкість обробки вхідних даних шляхом використання простіших інструкцій, які значно швидше виконуються процесором. Іншою перевагою є те, що науковці в галузі машинного навчання та обробки даних економлять свій час за рахунок зменшення порогу входження до розуміння внутрішньої структури алгоритму, а також витрачають менше на повторні багатократні запуски того ж алгоритму на різних наборах даних. Аналогічно дана перевага проявляє себе і для звичайних користувачів, що витрачають менше часу на очікування під час запуску даного алгоритму на великого розміру вхідних даних.

Апробація роботи. Основні положення і результати роботи були представлені та обговорювались на IX науковій конференції магістрантів та аспірантів ”Прикладна математика та комп’ютинг”ПМК-2017 (Київ, 19–21 квітня 2017 р.) та опубліковані у збірнику тез за результатами конференції; збір даних та їх попередня обробка, а також розміщення проміжних результатів було здійснено на веб-ресурсі kpidata.org (жовтень 2015 - до сьогоднішнього дня); доступ до вхідних даних опитування розміщено для вільного користування в онлайн-режимі github.com/kpidata/datasets.

Структура та обсяг роботи. Магістерська дисертація складається з вступу, п’яти розділів, висновків та додатків.

У вступі надано загальну характеристику роботи, виконано оцінку поточного

стану проблеми, обґрунтовано актуальність напрямку досліджень.

У першому розділі розглянуто теоретичні відомості, існуючі алгоритми класифікації текстових даних, наведено математичні основи, що використовуються для побудови моделей. Розглянуті загальні підходи до автоматизованої класифікації текстових даних та поширені алгоритми, що застосовуються в даній галузі. Основну увагу приділено всьому процесу обробки даних: від їх початкового збору до безпосереднього застосування прогностичної моделі.

У другому розділі здійснено аналіз існуючих алгоритів, проведено дослідження їх внутрішньої реалізації та математичного апарату, що лежить в їх основі. Були розглянути переваги і недоліки кожного класу алгоритмів та окреслена область їх застосування. Було виділено основні вимоги до розроблюваного алгоритму та обрано конкретні шляхи оптимізації, які будуть використані під час його розробки та покращення. Визначено вплив даних змін на результуючу модель та обґрунтовано доцільність здійснення даних модифікацій.

У третьому розділі запропоновано засоби реалізації для кожного з етапів методу; наведено огляд архітектурних підходів до організації програмного забезпечення; обґрунтувано вибір мікросервісної архітектури; запропоновано структуру та особливості реалізації кожного з мікросервісів, наведено відповідні графічні матеріали, що ілюструють взаємодію елементів системи.

У четвертому розділі наведено результати роботи алгоритму, підтверджено на практиці гіпотезу про те, що застосування розробленого алгоритму надає виграш у швидкодії; отримано підтвердження того, що використання однорідних інструкцій дозволяє зменшити витрати ресурсів процесора; здійснено порівняння точності та швидкості роботи з існуючими алгоритмами; зроблено висновок щодо можливості застосування даного підходу для використання з різними алгоритмами та вхідними даними для вирішення задачі класифікації; запропоновано шляхи покращення та вектори розвитку для подальшої роботи.

У п’ятому розділі подано аналіз програмного продукту, його оцінку та перспективи для виходу на ринок. Наведені слабкі та сильні сторони проекту, порівня-

ння з аналогами та конкурентоспроможність. Проведено оцінку розміру необхідних інвестицій, обсягу ресурсів, що потрібно залучити та показників прибутку за умови подальшої комерціалізації проекту.

У висновках проаналізовано отримані результати роботи.

У додатках наведено фрагменти програмної реалізації запропонованого способу та копії графічних матеріалів.

Робота виконана на 80 аркушах, містить 2 додатки та посилання на список використаних літературних джерел з 30 найменувань. У роботі наведено 14 рисунків та 4 таблиці.

Ключові слова: класифікація, прогностичні моделі, апроксимація моделі, датасет, машинне навчання.