НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем

“На правах рукопису”
УДК 004.021 045420

«До захисту допущено»
Науковий керівник
кафедри

_________ І. А. Дичка
(підпис)

“___” ___________ 2017 р.

Магістерська дисертація

зі спеціальності 8.05010302 “Інженерія програмного забезпечення”

на тему: МЕТОД АВТОМАТИЗОВАНОЇ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ДАНИХ НА ОСНОВІ ГІБРИДНИХ МОДЕЛЕЙ

Виконав: студент 6 курсу, групи КП-52м
Бегерський Михайло Васильович ___________
(підпис)

Науковий керівник доцент., к.т.н., доцент Заболотня Т.М. ___________
(підпис)

Рецензент доцент, к.т.н., доцент Дідковська М.В. ___________
(підпис)

Рецензент доцент, к.т.н., доцент Марченко О.І. ___________
(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____________
(підпис)

Київ – 2017

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, порівняння 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 92 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.

ABSTRACT

Relevance of the research We encounter significant increase of data in the global network and processing of that data becoming more and more difficult. There are a lot of machine learning and data classifications algorithms that come to the rescue and help automate manual work and analysis. Amount of users who want to have easier access to that data in unified format is also increases. That is why algorithms for data classification and clusterization have gain their growth. Speaking from the point of data scientiests these algorithms are not the best option in every use case and need to be tuned in order to achieve the best results. Therefore universal methods that will be able to automatically classify various input data have such a high demand on the market. Moreover those algorithms are required to have intuitive implementation not only for the creator of the algorithm but for the users and regular developers too.

Object of research is a process of creating an algorithm to build universal predictive model.

Subject of research is a set of methods for building predictive models and algorithms of data classification.

Project's goal: developing a new algorithm for predictive model creation that will show the same accuracy as competitors and have simple implementation.

Methods of research. There are methods of data mining, data classification techniques and statistical methods in current project.

Scientific innovation of the projects includes:

1. Universality of a model that allows to eliminate differences between internal implementations of algorithms and use it with the same data not losing the accuracy among with better performance results.

2. Process of creation for such a model was developed and generialized for any arbitrary one.

3. Performance of the model developed was confirmed.

On the current state of the projects this approach allows usage of a model on real- world data and might be used as drop-in replacement for existing algorithms.

Practical value of the projects is based on the significant increase of the performance when using the model created by the method described. Using model approximation and similar set of cpu instructions can outcome in performance boost. The other advantage allows to decrease average time for data scientiests to spend while examining internal structure of the models and details of algorithm's implementation. Model shows great results on a large datasets too and therefore it will decrease total amount of time when launching multiple times or using it with large data.

Project's approbation. Main ideas and summary for the projects were presented and discussed during The IX annual scientific conference "Applied math and computing" and has been published in corresponding set of theses. Data mining and its preprocessing as well as publishing of the results has been made on the web-resource kpidata.org; access to the input data is free and data is stored under version control system on the GitHub (github.com/kpidata/datasets).

Project's structure. Master's dissertation consists of introduction, five sections, summary and appendix.

Introduction shows general overview of the project, description of the cur- rent state for the problem and shows the relevance of such a topic of research.

First section describes theoretical background, current set of algorightms that are used for text classification, mathematical background for those methods and algorithms for predictive models creation. General approaches that are used for text classification have been overviewed. Main point of the section is a process of data mining starting from initial collection of data and resulting in usage of predictive model.

Second section contains analysis of existing algorithms and research of their 2 internal implementation was made. Advantages and disadvantages of each class of algorithms was showed and fields for their application were described. List of requirements for the resulting algorithm was created and then suggested a ways for both optimization and improvement. Influence of the changes on the resulting model was inspected and then necessity of making these changes was confirmed.

Third section shows implementation of every stage for the method described; architecture and design solution for the resulting software are discussed; usage of microservices approach is justified; structure and internal implementation of each microservice is provided as well as charts and diagrams that shows inter- action between components of the system.

Forth section provides results of algorithm usage and significant performance increase was confirmed. Confirmation for the advantages of usage of homogeneous cpu instructions has been made. The comparison of accuracy and performance for the developed algorithm and its competitors was conducted. Ways for the future development and improvement were suggested.

In fifth section analysis of software is provided as well as general estimation and possibilities to join a market. Strong and weak sides of the project are high- lighted and comparison with competitors is made. Estimation of investments amount and the amount of resources that need to be invoked for the successful results has been made.

Summary briefly overviews achieved results and highlights key features of the work done.

Appendix consists of significant source code snippets and code for the main modules.

The project consists of 92 pages, has 1 additional code listing, references with 20 entries, 14 figures and 24 tables.

Keywords: classification, predictive models, model approximation, dataset, machine learning.

СПИСОК ВИКОРИСТАНИХ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

1. Text mining. Классификация текста. Пример классификации документов с использованием программных алгоритмов statistica. Режим доступу: URL http://statosphere.ru/blog/135-text-mining1.html.
2. E. Bauer and R. Kohavi. *An empirical comparison of voting classification algorithms: Bagging, boosting and variants*. 1999.
3. Y. Bengio. *Learning deep architectures for AI.* 2009.
4. Hinrich Schütze Christopher D. Manning, Prabhakar Raghavan. *An Introduction to Information Retrieval.* Cambridge University Press., 2009.
5. John Doe. *The Book without Title*. Dummy Publisher, 2010.
6. P. Domingos and M. Pazzani. *On the optimality of the simple Bayesian classifier under zero-one loss*. 1997.
7. Robert W. Fairlie. *Kauffman Index of Entrepreneurial Activity*. Kansas City: Ewing Marion Kauffman Foundation, 2014.
8. Brad Feld. *Startup communities: Building an entrepreneurial ecosystem in your city*. Hoboken, NJ: John Wiley & Sons, 2012.
9. Steven Finlay. *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st ed.).* Basingstoke: Palgrave Macmillan., 2014.
10. Micheline Kamber Han Jiawei and Jian Pei. *Data mining: concepts and techniques.* Morgan Kaufmann., 2006.
11. G.Hultenand P.Domingos. *Mining complex models from arbitrarily large databases in constant time.* Edmonton, Canada, ACM Press, 2002.
12. E.Frank I.Wittenand M.Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Mateo, CA, 2011.
13. Helland I.S. *Steps Towards a Unified Basis for Scientific Models and Methods*. World Scientific., 2010.
14. Stapleton J.H. *Models for Probability and Statistical Inference.* Wiley- Interscience.,

2007.

15. Young-Hoon Kwak. *A brief history of Project Management*. Greenwood Publishing Group, 2005.
16. J. S. Long. *Regression Models for Categorical and Limited Dependent Variables (1st ed.)*. Sage Publications, Inc., 1997.
17. T. M. Mitchell. *Machine Learning.* McGraw-Hill, New York, NY, 1997.
18. J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK, 2000.
19. Arthur Samuel. *Some Studies in Machine Learning Using the Game of Checkers.* IBM Journal of Research and Development, Volume 44, 1959.
20. F. Sebastiani. *Machine Learning in Automated Text Categorization*.
21. Dane Stangler. *The Economic Future just Happened*. Kansas City: Ewing Marion Kauffman Foundation., 2009.
22. Martin Stevens. *Project Management Pathways*. Association for Project Management. APM Publishing Limited, 2002.
23. Sholom M. Weiss and Nitin Indurkhya. *Predictive Data Mining.* Morgan Kaufmann., 1998.
24. RandR.Wilcox.*Fundamentals of Modern Statistical Methods.* NewYork: Springer, 2010.
25. Шмидт С. Бирман Г. *Капиталовложения. Экономический анализ инвестиционных проектов.* М.: ЮНИТИ-ДАНА, 2003.
26. Лапыгин Ю. Н. *Управление проектами: от планирования до оценки эффективности*. М.: Омега-Л, 2008.