ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра комп'ютерних систем та технологій

(повна назва кафедри (предметної, циклової комісії))

# Дипломна робота

на здобуття ступеня вищої освіти «магістр»

(освітньо-кваліфікаційний рівень)

на тему: «Методи класифікації незбалансованих даних»

«Unbalance Data Classification Methods»


Виконав: студент денної форми навчання

спеціальності 123 – Комп'ютерна інженерія .

(шифр і назва напряму підготовки, спеціальності)

Маслєєв Олексій Костянтинович

(прізвище, ім'я, по-батькові)

Керівник д-р техн. наук, доцент. Гунченко Ю.О.

(науковий ступінь, вчене звання, прізвище та ініціали, **підпис**)

Рецензент д-р техн. наук, проф. Малахов Є.В

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент кандидат фіз.мат. наук доцент Розум М.В.

(науковий ступінь, вчене звання, прізвище та ініціали)


Рекомендовано до захисту:

Протокол засідання кафедри

№_____ від «___»_____ 2019 р.

Завідувач кафедри

_____ Ю.О. Гунченко

(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК № ___

протокол № __ від «___» ____ 2019 р.

Оцінка_____/_____/_____

(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

_____ О.О. Арсірій

(підпис) (прізвище, ініціали)

Одеса – 2019

# АННОТАЦІЯ

У сучасному світі існує велика кількість інформації, і її обсяг з кожним днем зростає. Але деякі дані можуть завдавати шкоди. Дуже важливо визначати такі дані вчасно і точно, щоб уникнути отримання збитку від них.

Існує досить багато алгоритмів, які дозволяють проводити класифікацію даних, однак не всі мають достатньо високу точність. Основною проблемою для всіх алгоритмів є наявність незбалансованих даних, коли об'єктів одного типу в багато разів більше ніж інших.

Метою даної роботи є підвищення якості класифікації об’єктів з незбалансованих даних. В роботі вивчаються алгоритми та різні допоміжні методи які дозволяють проводити класифікацію з високою точністю.

У даній дипломній роботі розроблено метод, який дозволяє підвищити точність класифікації при роботі з незбалансованими даними. Були проаналізовані існуючі алгоритми класифікації та на їх основі розроблено новий метод. Отриманий метод має перевагу в точності в порівнянні зі звичайними методами що дуже важливо оскільки всі дослідження проводилися на незбалансованих даних.

# АННОТАЦИЯ

В современном мире существует большое количество информации, и её объем растет с каждым днем. Но некоторые данные могут наносить ущерб. Очень важно определять такие данные вовремя и точно во избежание получения ущерба от них.

Существует достаточно много алгоритмов, которые позволяют проводить классификацию данных, однако не все имеют достаточно высокую точность. Основной проблемой для всех алгоритмов является наличие несбалансированных данных, когда объектов одного типа во много раз больше чем других.

Целью данной работы является повышение точности классификации объектов из несбалансированных данных. В работе изучаются алгоритмы и различные вспомогательные методы которые позволяют проводить классификацию с высокой точностью.

В данной дипломной работе разработан метод, который позволяет повысить точность классификации при работе с несбалансированными данными. Были проанализированы существующие алгоритмы классификации и на их основе разработан новый метод. Полученный метод имеет преимущество в точности по сравнению с обычными методами что очень важно поскольку все исследования проводились на несбалансированных данных.

# ABSTRACT

In the modern world exist a large amount of information and its volume is growing every day. However some data can be harmful. It is very important to identify such data on time and correctly to avoid damage from them.

There are many algorithms that allow to classify the data, however not all of them have a sufficiently high accuracy. The main problem for all algorithms is the presence of unbalanced data when objects of one type are many times more than others.

The purpose of this work is to improve the quality of the classification of objects with unbalanced data. Algorithms and various auxiliary methods are studied in the work which allow to carry out classification with high accuracy.

In this thesis work is developed a method that permit to increase classification accuracy when working with unbalanced data. Existing classification algorithms were analyzed and a new method was developed based on them. The obtained method has an advantage in accuracy compared to regular one which is extremely important since all investigations were carried out on unbalanced data.

ЗМІСТ

# ВСТУП

Сучасний світ не стоїть на місці и дуже швидко розвивається, з’являються велика кількість даних в різних сферах життя. Область застосування цих даних стає ширшою та виникає необхідність класифікувати ці данні оскільки не всі данні є безпечними. Але, з ростом кількості даних для аналізу виникає проблема їх класифікації, оскільки з’являється дисбаланс. Проблема дисбалансу з’являється коли у наборі даних один з класів має набагато менше екземплярів від іншого. Ця проблема часто зустрічається коли потрібно виявити шахрайство, виявлення аномалій, в медичній області, при необхідності встановити особистість особи та інше. В різних ситуація дисбаланс може буде дуже великим и дуже необхідно вчасно розпізнати дані яких небагато, оскільки саме ці дані можуть нести небезпеку.

Проблема дисбалансу зводиться до задачі класифікації даних що є розділом машинного навчання. При появі нових об’єктів, які невідомо до якого класу належать, потрібно вміти віднести їх до вже існуючого класу. Технології машинного навчання найбільш ефективно вирішує задачі класифікації.

Алгоритми що використовуються для рішення задачі класифікації потребують великої кількості обчислень, що раніше було проблемою, однак сьогодні продуктивність сучасної техніки зростає що надає можливість для якісного аналізу даних який складається з мільйонів записів.

Метою даної дипломної роботи є підвищення якості класифікації об’єктів з незбалансованих даних. Об’єктом є незбалансовані дані. Предметом є  алгоритми класифікації незбалансованих даних. В даній дипломній роботі аналізується існуючі алгоритми для класифікації даних та способи підвищення їх ефективності при роботі з незбалансованими даними на прикладі бінарної класифікації банківських транзакцій.

# ВИСНОВОК

В ході досліджень предметної області було розглянуто основні методи для класифікації даних. Було обрано базу з не збалансованих даних банківських транзакцій для роботи. Оскільки вибрані дані були незбалансованими було також розглянуто допоміжні методи для роботи з цими даними.

Було обрано та реалізовано три методи для класифікації, а саме метод опорних векторів, метод найближчого сусіда та метод Байєса. Для кожного методу було використано усі три допоміжних метода для роботи з незбалансованими даними. В результаті дослідження було знайдено найкращій з допоміжних методів. Результатом був високий рівень якості класифікації. Оцінка якості класифікації була зроблена за допомогою чотирьох мір, які показали не однозначність.

В ході спостережень було прийнято рішення про реалізацію додаткових методів для класифікації, які показали, що їх використання є цілком доцільним, оскільки вони покращили результати. Для кожного з методів класифікації було розроблено додатковий метод бегінгу. Розроблені методи можливо використовувати й для інших задач класифікації.

Було запропоновано та реалізовано алгоритм голосування з використанням усіх трьох класифікаторів який показ доцільність його використання. Результати свідчать про те що для покращення результатів цей алгоритм є найліпшим але з використанням допоміжних методів балансування даних. Такий комплексний метод дозволяє класифікувати об’єкти в режимі реального часу. Також зроблено висновки про те що дуже важливо обирати правильну кількість даних для навчання та тренування, та способи використання методів балансування, оскільки кількість обраних даних грає велику роль в якості роботи усіх методів та алгоритмів. При дотримані усіх вимог вдалось отримати дуже хороший результат.

# ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ


1. Mitchell T. Machine Learning. — McGraw-Hill Science/Engineering/Math, 1997.
2. Флах П. Машинное обучение. — М.: ДМК Пресс, 2015. — 400 с.
3. Bernhard Schölkopf, Alexander J. Smola Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. — MIT Press, Cambridge, MA, 2002
4. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.
5. Загоруйко Н.Г. Прикладные методы анализа данных и знаний — Новосибирск: Изд-во Ин-та математики, 1999. — 270 с.
6. Domingos P. On the optimality of the simple Bayesian classifier under zero one loss / P. Domingos, P. Pazzani. // Machine Learning. – 1997.
7. Шеремет О. Метод опорних векторів / О. Шеремет, В. Садовой. // Мат. Мод. № 1. – 2013.
8. Breiman L. Random forests // Machine learning. — 2001.
9. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. — 2011.
10. Burges C.J.C. A tutorial on support vector machines for pattern recognition.// Data Mining and Knowledge Discovery, — 1998.
11. Quinlan J.R. C4.5 Programs for machine learning. —San Mateo, 1993.
12. База даних банківських транзакцій − [Электронный ресурс] / Режим доступа: https://www.kaggle.com/mlg-ulb/creditcardfraud - 09.09.2019
13. В.Г. Козачков, О.К. Маслєєв, Ю.О. Гунченко /Використання Нейронних Мереж у Системах Розпізнання Мови// Шістнадцята всеукраїнська конференція студентів і молодих науковців «Інформаційні системи та технології ІСТ-2019» с.270-272- Одеса 2019.