ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
КАФЕДРА ІНФОРМАТИКИ І КОМП'ЮТЕРНИХ НАУК

Пояснювальна записка

до дипломної бакалаврської роботи

на тему:

ЗАСТОСУВАННЯ АЛГОРИТМУ НЕЧІТКОГО ЛІСУ ДЛЯ РОЗВ'ЗАННЯ ЗАДАЧІ РЕГРЕСІЇ МОЛЕКУЛЯРНИХ ДАНИХ З ПОПЕРЕДНІМ ВІДБОРОМ НЕЗАЛЕЖНИХ ЗМІННИХ

Виконав: студент 4 курсу, групи 4КН
спеціальності 122 «Комп'ютерні науки»

Єрошенко В.П.

Керівник: Литвиненко В.І.

Рецензент: Г.В. Рудакова

Херсон – 2021 р.

Факультет	**Кібернетики та системної інженерії**
Кафедра	**Інформаційних технологій та дизайну**
Рівень вищої освіти	**перший (бакалаврський) рівень**
Галузь підготовки	**12 «Інформаційні технології»**
	(шифр і назва)
Освітньо-професійна програма	**Комп'ютерні науки**
	(назва)
Спеціальність	**122 «Комп'ютерні науки»**
	(шифр і назва)

ЗАТВЕРДЖУЮ
Завідувач кафедри ІКН,
професор
_____В.І. Литвиненко
«____» _____ 2021 року

З А В Д А Н Н Я
НА ДИПЛОМНУ РОБОТУ СТУДЕНТА

Єрошенко Владислав Павлович
(прізвище, ім'я, по батькові)

1. Тема роботи: Застосування алгоритму нечіткого лісу для розв'язання задачі регресії молекулярних даних з попереднім відбором незалежних змінних керівник роботи Литвиненко Володимир Іванович, завідувач кафедри інформатики і комп'ютерних наук, професор
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом ХНТУ від 26.11. 2020 р. № 644-с
2. Строк подання студентом роботи
04.06.2021
3. Вихідні дані до роботи _____

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): Вступ. Розділ 1. Аналіз алгоритмів нечіткого лісу та випадкового лісу. Заходи змінної важливості та алгоритм нечітких лісів. Розділ 2. Кластеризація. Особливості meanshift Розділ 3.Визначення Ентропії. .Її проблеми та вирішення їх.Різновиди ентропії. Розділ 4. .Основні поняття.різновиди FuzzyForest.Характеристика Rstudio. Розділ 5. Практична частина. Створення програмного продукту на основі RandomForest. Розділ 6. Охорона праці (розробка програмного забезпечення)

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

Таблиць – 4,

Формул – 7,

Рисунків –16 ,

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
6	к.с.-г.н., доцент Малєєв В.А.		

7. Дата видачі завдання 08.02.2021

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	Примітка
1	Здійснення аналітичного огляду літературних джерел та аналіз сучасних методів програмного забезпечення	10.02.2021-01.03.2021	
2	Аналіз проблеми підбору регресії у «нечітких лісах»	02.03.2021-15.03.2021	
3	Аналіз предметної області - регресія , як спосіб вирішення молекулярних задач	16.03.2021-01.04.2021	
4	Аналіз особливостей та характеристики мови програмування R	02.04.2021-10.04.2021	
5	Аналіз програмного забезпечення RandomForest в середовищі розробки R.	11.04.2021-21.04.2021	
6	Створення основних етапів розробки регресії FuzzyForest. Опис основних бібліотек і модулів для написання коду.	01.05.2021-10.05.2021	
7	Тестування та відладка програми щодо регресії, виправлення помилок в програмному коді.	11.05.2021-21.05.2021	
8	Опис основних вимог охорони праці під час роботи програміста. Опис робочого місця програміста. Аналіз організації робочого місця програміста.	22.05.2021-04.06.2021	

Студент _____Єрошенко В.П.

(підпис) (прізвище та ініціали)

Керівник роботи _____ Литвиненко В.І.

(підпис) (прізвище та ініціали)

ЗМІСТ

РЕФЕРАТ **11**

ABSTRACT **11**

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ **12**

ВСТУП **13**

РОЗДІЛ 1. Аналіз алгоритмів нечіткого лісу та випадкового лісу. Заходи змінної важливості та алгоритм нечітких лісів **16**

1.1 Заходи змінної важливості 16

1.2 Мульти-класифікаційні методи. Комбінаційні методи 17

1.3 Вимірювання важливості змінних 20

1.4 Значення Gini 21

1.5 Значення функції RandomForest 22

1.6 Випадкова близькість лісу 23

Висновок до розділу **24**

Розділ 2. Кластеризація. Особливості meanshift **26**

2.1 Кластеризація 26

2.2 Алгоритм MeanShift 27

2.3 Постановка задачі та алгоритми середнього зсуву 30

2.4 Два основних типи алгоритмів середнього зсуву 32

2.4 Кластеризація шляхом розмиття середнього зсуву 32

2.6 Подібність та відмінності між MS та BMS 33

2.7 Вибір смуги пропускання 34

2.8 Вибір ядра 35

2.9 Матричне формулювання BMS та узагальнення BMS 37

2.9 Складання нових точок у кластери 38

2.10 Переваги та недоліки алгоритмів середнього зсуву 38

2.11 Витоки алгоритму середнього зсуву 40

2.12 Зближення Bayesian Model Averaging software 43

Висновки до розділу **51**

Розділ 3. Визначення Ентропії.Різновиди ентропії. **52**

3.1 Постановка проблеми та її складність 52

3.2 Процес відбору функцій 55

3.3 Методи обгортання 57
3.4 Методи фільтрування 59
3.5 Системна біологія 61
Висновок до розділу 63
РОЗДІЛ 4.Основні поняття.різновиди FuzzyForest.Характеристика Rstudio. 64
4.1 Основні поняття 64
4.2 Алгоритм FuzzyForest 65
4.3 Огляд WGCNA 67
4.4 Пакет FuzzyForest 69
4.5 Rstudio 71
Розділ 5.Практична частина. Створення програмного продукту на основі RandomForest. 74
5.1 Задання основних даних 74
5.2 Початок розподілу даних по структурі 75
5.3 Отримання результатів кластеризації та регресії 77
Висновок до розділу 79
РОЗДІЛ 6. ОХОРОНА ПРАЦІ (РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ) 80
6.1.Основні вимоги охорони праці під час роботи програміста 80
6.2.Опис робочого місця програміста 81
6.3.Розрахунок інформаційного навантаження програміста. 84
6.4 Організація робочого місця програміста 87
Висновки до розділу 89
ВИСНОВОК 90
Список використаних літературних джерел 91
ДОДАТКИ 93

РЕФЕРАТ

В данній дипломній роботі наведено огляд сучасних підходів до систем розрахунку за допомогаю випадкових лісів (RF) , за допомогаю яких ми можеме знаходити результати дуже тяжких формул , використовувати великий обсяг даних та розраховувати їх в наших цілях.

Розроблено програму для розрахунку великого обсягу даних на базі RF та RFR в середовищі програмування Rstudio на мові програмування R. Метод який був використаний називається – розрахунок даних за допомогаю випадкого лісу та регресії остаточних даних.

ABSTRACT

This thesis provides an overview of modern approaches to calculation systems using random forests (RF), with which we can find the results of very difficult formulas, use large amounts of data and calculate them for our purposes.

A program has been developed to calculate a large amount of data based on RF and RFR in the Rstudio programming environment in the R programming language. The method used is called - calculation of data using random forest and regression of final data.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

Скорочення, термін, позначення	Пояснення
ММВ	міри мінливості важливості
ВПРО	випадкові поглинання рекурсивних ознак
LASSO	Least Absolute Shrinkage and Selection Operator
RF	RandomForest
RFE	RandomForestEach
WGCNA	Weighted correlation network analysis
VIM	Visualization and Imputation of Missing Values
OOB	Out of bug
KDE	Ітерація середнього зсуву
CRAN	Comprehensive R Archive Network
BMS	Bayesian Model Averaging software
MS	Model for search
SVM	Support Vector Machine
ANN	Artificial Neural Network
RBF	Robust Backfitting
LPCM	Local Principal Curve Methods
ReLU	Rectified linear unit
SVRMHC	server for MHC-binding peptides
Bagging	**Bootstrap Aggregation**

ВСТУП

З часом розвитку людства все більше сфер яким раніше не приділялося належної уваги та часу на дослідження, через велику ресурсо витратність, зараз отримали можливість на реалізацію свого потенціалу, технічний прогрес може посприяти розвитку шляхом безпосереднього аналізу конкретної, значущою для дослідження проблеми в області медицини. Такі Проблеми вимагають іноваційного підходу до систематизації наявних данних, прогнозування наслідків можливих змін подальшого аналізу і застосування отриманих знань на практиці для забезпечення комфортних умов існування з поправкою на усі ситуації, які можна спрогнозувати.

Так як, дерева регресії – це легкі, але потужні моделі, які широко використовуються в різних областях застосування. Їх головна перевага залежить від простоти як на етапі навчання, так і на етапі виконнання. Крім того, РТ – це високо інтрепритовані моделі, тобто вони можуть бути використані для опису зв'язку між входами та виходом. РТ – це деревоподібний спрямований графік, в якому кожен внутрішній(не-листовий) вузол являє собою тест на атрибут, і кожен листовий вузол містить вихідне значення. Кожна гілка, яка складається з послідовності нелистових вузлів (тестів) та листового вузла (вихідне значення),

В епоху високопродуктивних технологій, таких як багатоколірна проточна цитометрія та наступне покоління, послідовність послідовностей генерації даних з високими розмірами стає все більш поширеною у біомедичних дослідженнях. Однак здатність генерувати дані значно перевершила нашу здатність аналізувати її. У біомедичних науках, а також в "полях Оміка" загальноприйнятим є те, що кількість параметрів набагато більша за кількість спостережень, так звана проблема $p << n$

. Ця проблема посилюється тим, що ознаки часто сильно корелюють і кореляційна структура часто апріорі невідома. Визначення важливих особливостей у цій ситуації було сферою інтенсивних досліджень у межах

спільнота статистики та машинного навчання. Хоча алгоритми вибору функцій, засновані на моделі, такі як LASSO можуть виявляти важливі особливості за наявності кореляції , це відбувається за рахунок формування параметричних припущень, які можуть не відповідати практиці. Якщо $p >> n$, система LASSO може вибрати не більше функцій. В умовах, коли існують групи високо корельованих ознак, LASSO має тенденцію довільно вибирати ознаки з групи та зводити інші до нуля. Випадкові ліси є популярним алгоритмом машинного навчання ансамблю. Випадкові ліси непараметричні, нелінійні, незручно паралелізуються, прості у впровадженні та описані як один з найкращих класифікаторів, що не продаються. Випадкові міри мінливості важливості лісу пропонують гнучку альтернативу вибірковим ритмам ознак на основі моделей. Хоча випадкові лісові ММВ продемонстрували здатність точно фіксувати справжню важливість функцій у налаштуваннях, де функції не залежать, загальновідомо, що випадкові лісові ММВ-файли є упередженими, коли функції співвідносяться з іншими. Нечіткі ліси обробляють корельовані риси, застосовуючи кусковий підхід. Спочатку ми оцінюємо мережеву структуру даних та розподіляємо набір функцій на різні модулі, так що кореляція в кожному модулі висока, а кореляція між модулями низька. У зв'язку з цим ми використовуємо функціонал Зважений аналіз мережі коекспресії генів, сувору основу для виявлення кореляційних мереж . Потім ми використовуємо випадкові поглинання рекурсивних ознак , щоб вибрати найважливіші функції з кожного модуля. Потім один остаточний ВПРО застосовується до решти функцій, вибираючи та класифікуючи найважливіші мелодії.

Актуальність теми. На сьогодні мільйони людей живуть не озираючись на проблеми здоров'я та свого співчуття. Таким чином люди не застережують себе від будь-яких пошкоджень , проблем та чи малого іншого , що може трапитися.

Сучасна біологія все частіше використовує методи машинного навчання

для широкомасштабного та складного аналізу біологічних даних. У галузі біоінформатики популярним вибором є техніка Random Forest (RF), яка включає ансамбль рішучих дерев та включає вибір функцій та взаємодії, природним чином у процесі навчання. Він непараметричний, інтерпретується, ефективний і має високу точність прогнозування для багатьох типів даних. Нещодавня робота з обчислювальної біології показала збільшення використання випадкового лісу завдяки його унікальним перевагам у роботі з малим обсягом вибірки, просторовим простором ознак та складними структурами даних

При використанні даного программного забезпечення, компанія зможе швидко, та що не маловажно, з точністю,спрогнозувати на основі накопичних данних певну ознаку проблеми , що має поліпшити їх роботу.

Мета і задачі дослідження. Дана робота присвячена розробці та створенню зручнії та функціональної програмі , що буде прогнозувати які є шанси у той чи іншій операції при знаходжені її мінімальних можливостях.

Апробація результатів бакалаврської кваліфікаційної роботи.

Єрошенко В.П Алгоритм нечіткого лісу для розв’язання задачі регресії молекулярних даних з попереднім відбором незалежних змінних.