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 класифікації молекулярних даних з попередньої редукцією незалежних змінних

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

Фещук А. О.

Керівник: Литвиненко В.І.

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

Фещук Андрій Олегович
(прізвище, ім'я, по батькові)

1. Тема роботи: Застосування алгоритму нечіткого лісу для класифікації молекулярних даних з попередньої редукцією незалежних змінних
керівник роботи Литвиненко Володимир Іванович, доктор технічних наук, професор
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом ХНТУ від _____ № ___
2. Строк подання студентом роботи
04.06.2020
3. Вихідні дані до роботи Матеріали та результати, отримані під час

проходження переддипломної практики, методичні вказівки, технічна література.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити): Вступ. Розділ 1. Аналітичний огляд літературних та інших даних. Розділ 2. Теоретична частина. Аналіз предметної. Розділ 3. Практична частина. опис створеного програмного рішення. Розділ 4. Охорона праці під час пандемії.

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

Таблиць –

Формул –

Рисунків –

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Розділ 4	к.с.н., доцент Малєєв В.О.		

7. Дата видачі завдання 10.02.2021

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 класифікації даних	02.03.2021-15.03.2021	
3	Аналіз предметної області - класифікації, як метод 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 програмне забезпечення RStudio. Дослідження Fuzzy forest	11.04.2021-21.04.2021	
6	Створення основних етапів розробки класифікації FuzzyForest. Опис основних бібліотек і модулів для написання коду.	01.05.2021-10.05.2021	
7	Тестування та відладка програми щодо класифікації, 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 _____Фещук А. О.

(підпис) (прізвище та ініціали)

Керівник роботи _____ Литвиненко В. І.

(підпис) (прізвище та ініціали)

ЗМІСТ

РЕФЕРАТ

В роботі наведено огляд сучасних підходів до класифікації молекулятних даних за допомогою алгоритму нечіткого лісу. Розглянуто етапи побудови fuzzy forest, методи попередньої обробки даних. Розроблено архітектуру обробки молекулярних даних із попередньою редукцією незалежних змінних. Для здійснення поставленної задачі використано декілька поширених методів обробки даних(нечітка кластеризація, зменшення ентропії, ранжування за допомогою fuzzy forest, та класифікація за допомогою випадкового лісу).

Створено програмну реалізацію в середовищі RStudio на основі функцій із пакетів мови програмування R. Проведено аналіз отриманих результатів.

ABSTRACT

The paper provides an overview of modern approaches to the classification of molecular data using the fuzzy forest algorithm. The stages of fuzzy forest construction, methods of data pre-processing are considered. The architecture of molecular data processing with preliminary reduction of independent variables is developed. Several common data processing methods were used to accomplish this task (fuzzy clustering, entropy reduction, fuzzy forest ranking, and random forest classification).

The software implementation in the RStudio environment on the basis of functions from packages of programming language R is created. The analysis of the received results is carried out.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

Скорочення, термін, позначення	Пояснення
VIM	Visualization and Imputation of Missing Values
WGCNA	Weighted correlation network analysis
RF	RandomForest

ВСТУП

Класифікація завжди була необхідна протягом усього існування людства. Ще для наших давніх предків, цей процес був необхідний для життя та розвитку. Сплеск інформації, що актуальній останні декілька десятків років що стала доступна компаніям та людям, ще більше посилюють цю проблему. Існує багато методів та алгоритмів, що вирішують питання класифікації. За останні рокі ми також спостерігаємо збільшення кількості підходів на основі систем класифікації, які, як було показано, дають кращі результати, ніж окремі класифікатори. Однак недосконала інформація неминуче з’являється в реалістичних сферах та ситуаціях. Помилки приладу або пошкодження через шум під час експериментів можуть призвести до появи інформації з неповними даними при вимірюванні конкретного атрибута.

Класифікація в області імунології, безумовно, стала необхідністю. Як випливає з назви, це процес класифікації даних. І для прийняття рішень потрібно прийняти багато даних. Часто це залежить від набору вхідних змінних. Класифікація залежить від ряду підтверджень та випадків даних. Протягом багатьох років «методи класифікації в інтелектуальному аналізі даних», «алгоритми класифікації в інтелектуальному аналізі даних» та «класифікація, що базується на правилах, у інтелектуальному аналізі даних» перетворилися на популярні теми. Тому тема моєї дипломної роботи настільки цікава та сучасна.

Методи класифікації з'єднань як потенційних ліків, які зв'язуються з конкретною ціллю, стають все більш важливими для розробки ліків. Для створення пристроїв класифікації необхідні навчальні набори ліків з відомою активністю. Для багатьох завдань такої класифікації доступна не тільки якісна, але і кількісна інформація про конкретний властивості (наприклад, про спорідненість зв'язування). Останній можна використовувати для побудови схеми класифікації для прогнозування цієї властивості для нових з'єднань.

Явна прогнозування складеного властивості зазвичай важче, ніж класифікація того, що властивість знаходиться нижче або вище заданого порогового значення. Отже, непряма класифікація, заснована на регресії, може привести до гірших результатів, ніж схема прямої класифікації. Фактично, спочатку дослідників цікавила тільки класифікація з'єднань як потенційних ліків.

Отримання точної інформації може бути надто дорогим або нежиттєздатним. Більше того, іноді може бути корисним використовувати додаткову інформацію від експерта, яка, як правило, подається через нечіткі поняття типу: маленька, більш-менш, близька до тощо. У більшості реальних проблем дані мають певний ступінь неточності. Іноді ця неточність досить мала, щоб її можна було безпечно ігнорувати. В інших випадках неточність даних можна змоделювати шляхом розподілу ймовірностей. Нарешті, існує третій тип проблем, коли неточність є суттєвою, і розподіл ймовірностей не є природною моделлю. Таким чином, існують певні практичні проблеми, коли дані за своєю суттю нечіткі.

Актуальність теми.

Проблеми молекулярної класифікації поширені в багатьох областях хімії, біохімії, фармації, медичної діагностики і інших додатків в сучасних науках про життя. У емпіричної регресії або завданню класифікації розглядаються об'єкти, які мають певні спільні відносинами, що характеризуються специфічними для об'єкта цільовими значеннями. Для молекул ліки ці цільові значення можуть ставитися до властивості, яке характеризує зв'язування. У задачі регресії ці цільові значення безперервно змінюються в певному інтервалі. Це може бути, наприклад, афінність зв'язування ліків, які зв'язуються з одним і тим же рецептором. Для завдання класифікації об'єкти будуть позначені на основі групою, до якої вони належать, з використанням дискретних цільових значень. Щоб вирішити задачу класифікації або регресії, нам необхідно співвіднести окремі об'єкти з їх відповідними цільовими значеніямі.t = + 1 / -1, який можна використовувати для характеристики об'єктів позитивного / негативного класу. Щоб створити

емпіричне пристрій (регресорів або класифікатор) для прогнозування невідомих цільових значень, потрібно навчальний набір пов'язаних об'єктів з відомими цільовими значеннями.

Можливість класифікації даних дозволяє швидко й точно розподілити їх на групи по окремих ознаках. Отже, стає необхідним включити обробку інформації з атрибутами, які, в свою чергу, можуть представляти відсутні та неточні значення як на етапах навчання, так і на етапі класифікації. Крім того, бажано, щоб такі методи були якомога надійнішими щодо шуму в даних. Така система знайшла б застосування як у науковій діяльності так і у бізнесі, що і визначає актуальність даної теми.

Мета і завдання дослідження.

Реалізувати алгоритм класифікації з попередньою редукцією незалежних змінних за допомогою можливостей представлених мовою програмування R.

Необхідно дослідити точність отриманих результатів та вивчити увесь процес формування структурі fuzzy forest.

Опрацювати обране середовище програмування та вдосконалити знання в області data-mining

В бакалаврській роботі поставлені та розв’язані ***наступні задачі***:

✓ Виконати всебічний аналіз об’єкта дослідження та предметної області;

✓ Дослідити поєднання інструментарію сучасних програмно – апаратних методів програмування із одних із методів класифікації;

✓ Використати на практиці знання отриманні при вивченні таких дисциплін, як «Data-mining», «Системне програмування», «Крос-платформне програмування», «Технологія створення програмних продуктів».

Об’єкт дослідження:

✓ Дерева рішень.

✓ Сучасні проблеми класифікації данних.

✓ Принципи побудови структури Fuzzy forest.

- ✓ Мова програмування R, її основні пакети.
- ✓ Програмне забезпечення RStudio.

Наукова новизна одержаних результатів полягає в наступному:

Отриманий функціонал, який буде здійснювати класифікацію молекулярних даних для полегшення роботи фахівцям та науковцям у сфері медицини та біології.

Класифікація даних - це популярна проблеми аналізу даних. Багато дослідників створили безліч інструментів для вирішення цих питань. Нечітка кластеризація, нечіткі дерева рішень та класифікатори ансамблів, такі як нечіткі ліси, є популярними інструментами, що використовуються для вирішення таких проблем. Необхідно описувати такі методи, щоб показати спосіб їх вирішення з проблемами класифікації даних.

Ми розглянемо проблему нечіткої кластеризації, яка є одним з найважливіших аспектів нечітких дерев, що базуються на кластерах.

Організації всього світу інвестують у ці алгоритми, щоб глибше засвоїти переваги та поведінку своїх клієнтів, отримати ширше представлення про роботу бізнес-процесів та опрацювати величезні об’єми інформації з досліджень у сфері біології, фізики та математики.

Класифікація має багато застосувань у сегментації споживачів, бізнес-моделюванні, маркетингу, кредитному аналізі та біомедичному та моделюванні реакції на наркотики.

З максимальною ефективністю поєднати інструментарій мови програмування R з можливостями робочої області для програмування RStudio. Реалізація структури рішень Fuzzy forest.

Біомедичні дослідження часто мають набагато більше предикторів, ніж спостереження. Ця проблема стає ще більш нерозв'язною, коли багато з параметрів мають високу кореляцію. Відсутність незалежності порушує основні припущення багатьох стандартних статистичних моделей. Крім того, багато біологічних систем включають складні моделі кореляції, включаючи взаємодії високого порядку та мережеві ефекти. Коли простір параметрів

великий, неможливо визначити взаємодії апріорі, і багато з цих взаємодій можуть бути невідомими. У багатьох геномних дослідженнях може бути лише невелика кількість змінних, які насправді важливі для фенотипу; тому може бути велика кількість параметрів шуму щодо важливих змінних.

Двома найбільш широко використовуваними алгоритмами машинного навчання в медичній науці є випадкові ліси (RF) та опорні векторні машини (SVM). Обидва мають високу точність прогнозування, але тут обговорюються лише RF. Популярність RF зумовлена їх відносною обчислювальною ефективністю, стійкістю до вибіжних показників та незмінністю до змішаних типів змінних. Більше того, RF порівняно важко переоцінити, вони непараметричні і, природно, обробляють взаємодії та змінні значення. RF були описані як найкращий "готовий" алгоритм, і вони користуються хорошою точністю прогнозування та широким використанням.

Методика класифікації, що використовується для упорядкування категорії різних наборів даних. Класи іноді називають цілями / мітками або категоріями. Основною метою моделювання аналітичної класифікації є схожість на завдання виконання (t) із змінної вхідної (x) для ізоляції вихідних змінних (y).

Метод найбільш звичайним методом є метод витримки. У цьому методі даний набір даних розділений на два розділи як тестові дані та дані тренувань 20% та 80% відповідно. Комплект тестових даних буде використаний для навчання моделі.