

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 математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота
на здобуття ступеня вищої освіти «магістр»
(освітньо-кваліфікаційний рівень)

на тему: Аналіз якості класифікації методами Data Mining
Analysis of the classification quality by Data Mining methods

Виконав: студент денної форми навчання
спеціальності 126 – Інформаційні системи та технології
(шифр і назва напрямку підготовки, спеціальності)

Нсуе Єнгонга Паскаль Хосеевич

(прізвище, ім'я, по-батькові)

Керівник к.фіз.-м.н. Петрушіна Т.І.

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

Рецензент к.фіз.-м.н. Шпінарева І.І.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент д.тех. н Князева Н.О.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:
Протокол засідання кафедри
№ _____ від «___» _____ 2019 р.

Завідувач кафедри

Захищено на засіданні ЕК № _____
протокол № _____ від «___» _____ 2019
р.

Оцінка _____ / _____ / _____
(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

(підпис) Є.В. Малахов
(прізвище, ініціали)

(підпис) О.О. Арсірій
(прізвище, ініціали)

Одеса - 2019

АННОТАЦИЯ

В дипломной работе разрабатывается тема «Анализ качества классификации методами Data Mining».

Цель работы – оптимизация выбора метода анализа данных в конкретной прикладной области путем оценки качества их применения.

Предмет исследования – методы прогнозирования и классификации результатов лабораторных исследований рака груди.

Объектом исследования являются методы классификации типа рака, отвечающие требованиям медицинской предметной области.

В качестве данных для обработки был взят датасет госпиталя Висконсинского университета собранный доктором Уильямом Х. Волбергом. На основе этих данных были получены оценки для 6 классификаторов и были выделены методы, наиболее удовлетворяющие критериям предметной области, в частности были выделены графические методы как наиболее легко интерпретируемые.

АННОТАЦІЯ

У дипломній роботі розробляється тема «Аналіз якості класифікації методами Data Mining».

Мета роботи – оптимізація вибору методу аналізу даних конкретної прикладної галузі шляхом оцінки якості їх застосування.

Предмет дослідження – методи прогнозування і класифікації результатів лабораторних досліджень раку грудей.

Об'єктом дослідження є методи класифікації типу раку, що відповідають вимогам медичної предметної області.

В якості даних для обробки був узятий датасет госпіталю Вісконсинського університету зібраний доктором Уільямом Х. Волберг. На основі цих даних були отримані оцінки для 6 класифікаторів і були виділені методи, які найбільш задовольняють критеріям предметної галузі, зокрема були виділені графічні методи як ті що найбільш легко інтерпретуються.

ABSTRACT

In the thesis, the theme “Analysis of the classification quality by Data Mining methods” is developed.

The purpose of the work is to optimize the choice of analysis methods in a particular industry by assessing the quality of their applications.

Subject of research – methods for predicting and classifying breast cancer.

The object of the study is a cancer type classifier that meets the requirements of the medical subject area.

The data set for processing was taken by the data set of the University of Wisconsin hospital collected by Dr. William H. Volberg. Based on these data, estimates were obtained for 6 classifiers and methods were selected that most satisfy the criteria of the subject area, in particular, graphic methods were identified as the most easily interpreted.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	6
1 ОПИСАНИЕ ИСХОДНЫХ ДАННЫХ. МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ	8
1.1 Перспективы использования методов машинного обучения в медицине	8
1.2 Выбор языка разработки.....	10
1.3 Анализ библиотек языка Python и описание методов машинного обучения.....	11
1.3.1 Дерево решений	13
1.3.2 Случайный лес.....	15
1.3.3 Градиентный бустинг	17
1.3.4 Логистическая регрессия.....	18
1.3.5 K ближайших соседей.....	20
1.3.6 Многослойный перцептрон	21
Вывод	22
2 ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ.....	24
2.1 Первичный анализ и предобработка исходных данных	24
Вывод	26
2.2 Применение методов классификации, отбора признаков	26
Вывод.	29
3 АПРОБИРОВАНИЕ РЕЗУЛЬТАТОВ	30
3.1 Анализ и интерпретация полученных результатов	30
Accuracy.....	30
Precision	30
Вывод.	34
3.2 Практическое обоснование результатов исследования	34
Вывод.	35
ЗАКЛЮЧЕНИЕ	36
ПЕРЕЧЕНЬ ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ.....	37
ПРИЛОЖЕНИЯ А МАТРИЦА КОРРЕЛЯЦИИ	39
ПРИЛОЖЕНИЯ Б КОД ПРОГРАММЫ	40

ВВЕДЕНИЕ

Анализ и сравнение информации является одной из самых важных задач в любой отрасли человеческой деятельности, в том числе в научных и медицинских исследованиях. Почти вся оцифрованная информация представлена в виде текстов, изображений или таблиц. Прямая работа с текстами занимает порядка 80% времени, которое тратят специалисты на обработку и сравнение информации для принятия решений.

Машинное обучение возникло в результате разделения науки о нейронных сетях в рамках науки об искусственном интеллекте на методы обучения сетей и типов топологий сетевой архитектуры, включающих некоторые другие области, такие как методы математической статистики и теория дискретного анализа. Классические методы машинного обучения, используемые в задачах классификации и сравнения данных, такие как байесовские методы, нейронные сети, векторные методы, относятся к группе методов обучения с учителем. Для создания классификатора эти методы используют наборы данных, объекты которых уже распределены по классам.

В процессе обучения формируются оптимальные параметры классификатора, приводящие к наименьшей ошибке классификации. Использование методов классификации с учителем очень эффективно при наличии репрезентативного набора данных. Однако часто бывает трудно или невозможно получить такой набор на практике. Проблемы возникают из-за недостаточного количества имеющихся примеров, наличия ошибок в данных, несоответствия между распределением обучающей и тестовой выборок. Таким образом, из-за трудностей с формированием обучающей выборки качественного обучения и точность распознавания методов обучения с учителем значительно ухудшается.

Основные задачи, решаемые с помощью методов машинного обучения:

- классификация – выполняется с помощью обучения с учителем;
- регрессия – выполняется с помощью обучения с учителем на этапе тестирования, является частным случаем задач прогнозирования;
- кластеризация – выполняется с помощью обучения без учителя;
- понижение размерности данных и их визуализация – выполняется с помощью обучения без учителя.

Целью дипломной работы является оптимизация выбора метода анализа данных в конкретной прикладной области путем оценки качества их применения.

Для достижения поставленной цели необходимо решить следующие задачи:

- сформулировать требования к создаваемой системы;
- провести фильтрацию и факторизацию изначальных данных;
- разделить выборку на тестовую и обучающую;
- выбрать инструментальная среда разработки и выполнить программную реализацию классификатор.
- разработать критерии оценки сравнения алгоритмов;
- провести их сравнения;

Предмет исследования – методы прогнозирования и классификации результатов лабораторных исследований рака груди.

Объектом исследования являются методы классификации типа рака, отвечающие требованиям медицинской предметной области.

ЗАКЛЮЧЕНИЕ

Исследования в области применения методов машинного обучения к диагностике заболеваний рака груди являются актуальными на сегодняшний день и проводятся как в Украине, так и за рубежом.

В рамках данной работы были проведена подготовка медицинских данных к автоматической обработке, а также разработана программа, извлекающая необходимую информацию из них в автоматическом режиме.

Были реализованы некоторые методы анализа медицинских данных проведена апробация этих методов на подготовленных реальных данных. Полученные результаты были использованы для анализа качества применения различных методов машинного обучения. Были предложены подходы к повышению качества методов в указанной предметной области.

По результатам исследований метод к-ближайших соседей показал большую по сравнению с остальными методами адекватность и предсказательную способность по всем выбранным критериям. Метод рекурсивного отбора признаков в целом значительно не улучшил и не ухудшил качество классификаторов.

Результаты работы показали обоснованность развития данной проблематики. Дальнейшие исследования в области применения на практике методов машинного обучения к диагностике заболеваний рака груди могут помочь врачу не только в извлечении информации из массива данных показаний диагностической аппаратуры и анализе этих данных, но и в сокращении времени постановки диагноза и увеличении его точности.

В данной работе были проанализированы характеристики методов классификации, показана их сходимость для данной задачи. Показано что применение разных методов датамайнинга приводят к похожим результатов для наших данных, следовательно, доказана успешная применимость этих методов.

Сделаны выводы, что возможно перейти от менее точных, но хорошо визуализируемых методов, к методам более точным и близким по результатам.

ПЕРЕЧЕНЬ ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Applications of Machine Learning in Cancer Prediction and Prognosis Joseph A. Cruz, David S. Wishart
2. Gene expression based cancer classification Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman
3. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / А.Н. Горбань // Сиб. журн. вычисл. математики. – 1998. – Т. 1, № 1. – С. 11–24.
4. Boughorbel S. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric / S. Boughorbel, F. Jarray, M. El-Anbari // PLoS ONE 12(6). – 2017. – 17 p.
5. Documentation of scikit-learn 0.19.1 [Электронный ресурс]. – URL: <http://scikit-learn.org/stable/documentation.html>
6. Воронцов К.В. Лекции по логическим алгоритмам классификации [Электронный ресурс]. – URL: <http://www.ccas.ru/voron/download/LogicAlgs.pdf>
7. Мюллер А., Гвидо С., Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными, Москва 2017, 480 с
8. Евклидово расстояние [Электронный ресурс]. URL: <http://statistica.ru/glossary/general/evklidovo-rasstoyanie/> (дата обращения: 16.05.2018).
9. Воронцов К.В., Лекции по метрическим алгоритмам классификации. 2008, 16с.
10. Воронцов К.В. Лекции по алгоритмическим композициям. – 2007. – 45с. [Электронный ресурс]. – URL: <http://www.ccas.ru/voron/download/Composition.pdf>
11. Breiman L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – P. 123–140

12. Second Edition, Brett Lantz, Machine Learning with R// Packt Publishing, 2013, 454 p.
13. База данных [Электронный ресурс] <https://github.com/jeffheaton/aifh/blob/master/vol1/python-examples/datasets/breast-cancer-wisconsin.csv>
14. Приложение [Электронный ресурс] <https://colab.research.google.com/drive/10wMZq0GV6FZAolmuPulPs9is69NL0ddw>