Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники»

УДК 004.891.2, 004.42

Пилипенко
Виталий Викторович

Диагностическая модель консультационной медицинской системы

АВТОРЕФЕРАТ

на соискание степени магистра инженерных наук
по специальности 1-40 80 02 Системный анализ, управление и обработка информации

Научный руководитель
Герман Олег Витольдович
кандидат технических
наук, доцент

Минск 2022 г.

КРАТКОЕ ВВЕДЕНИЕ

Диагностическая модель консультационной медицинской системы представляет собой комплекс программных средств, которые будут полезны при диагностике отдельных заболеваний.

В качестве объекта настоящей диссертационной работы выступает алгоритм классифицирующих деревьев для принятия диагностических решений (decision trees), а также случайные леса (random forests) в качестве их расширения (улучшения). На основе этих механизмов в диссертации разработаны диагностические модели для применения в задачах медицинской диагностики. В качестве примера диагностируемого заболевания выступает весьма распространённое в наше время заболевание – сахарный диабет.

Актуальность темы определяется значительно возросшими возможностями анализа данных, машинного обучения и компьютерных вычислений в области Data Mining для реализации систем, предоставляющих результат в режиме реального времени. Data science проникает во все аспекты нашей жизни. Такая популярность существенно обусловливается расширением доступности интернета и развитием веба.

В диссертационной работе разработан веб-сайт с удобным, доходчивым графическим интерфейсом, который можно развернуть на каком-либо сервере (например, облачном), доступном из любой точки мира. Разработанное веб приложение использует довольно сложные вычисления на основе средств Python для имплементации алгоритма random forest и classifying tree. В работе представлены результаты проведенных исследований в системах аналитического программирования Python и R, выполнен их сравнительный анализ.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования. Целью диссертационной работы является изучение и использование механизма классифицирующих деревьев (случайных лесов) для принятия решений в системах on-line медицинской диагностики.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Изучить модели и возможности классифицирующих деревьев (случайных лесов).

2. Разработать модели диагностики диабетического заболевания на основе механизмов классифицирующих деревьев (случайных лесов).

3. Выполнить программную имплементацию.

4. Провести анализ исследований.

Новизна полученных результатов:

- Проведена сравнительная характеристика средств (библиотек, пакетов, функций, параметров и т.д.) для работы как с классифицирующими деревьями принятия решений, так и со случайными лесами в R и Python; установлено влияние параметров на точность моделей, наличие/отсутствие/удобство конфигурации параметров в обоих языках.
- Выделены наиболее важные для расщепления деревьев признаки (features) на примере открытого набора данных для диабета, доказано отсутствие прямой зависимости между весом (важностью) признака и точностью модели.
- Рассчитаны средневзвешенные величины точности (вероятности) для каждого типа модели, подтверждена высокая эффективность лесов, их способность делать более точные предсказания даже при отсутствии весомого признака (feature).

Положения, выносимые на защиту.

1. Модели машинного обучения для decision trees и random forest; анализ их параметров для достижения максимальной точности, их точность и средневзвешенные значения на основе важности признаков (features);

2. Сравнительная характеристика средств языков R и Python для построения, настройки, обучения, визуализации моделей посредством алгоритмов классифицирующих деревьев и случайных лесов;

3. Веб-приложение (сайт), обладающее доходчивым интерфейсом (user-friendly UI) для людей, не владеющих специальными знаниями ни в области медицины, ни науки о данных, использующее созданный на основе обученной посредством алгоритмов машинного обучения модели.

Апробация результатов диссертации. Результаты магистерской диссертации были представлены на конференции «Информационные технологии и управление 2021» в секции «Автоматизированные системы обработки информации».

Структура и объём диссертации. Данная магистерская работа обладает следующей структурой:

1. Введение

2. Анализ проблемы и существующих средств решения. Обзор классифицирующих деревьев

3. Алгоритмическое описание математической модели. Разработка программных средств и проведение эксперимента на Python

4. Разработка программных средств и проведение эксперимента на R. Сравнительная характеристика имплементации на Python и R

5. Пользовательский интерфейс

6. Заключение

7. Список источников

Полный объём диссертации: 90 страниц.

Количество изображений: 58.

Количество использованных источников: 31.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Анализ проблемы и существующих средств решения. Обзор классифицирующих деревьев. Описывается проблема классификации в науке о данных, в частности, в целях медицинской диагностики, перечисляются методы, используемые для принятия решений. Приводится краткая характеристика деревьев принятия решений, а также указываются преимущества использования случайных лесов в качестве альтернативы классическим деревьям. Даются определения основным понятиям и затрагиваются проблемы, которые возникают в процессе обучения моделей, например, переобучение (overfitting).

Глава 2. Алгоритмическое описание математической модели. Разработка программных средств и проведение эксперимента на Python. В главе сообщаются результаты исследования алгоритма смешанных лесов и классических классифицирующих деревьев для принятия решений, проводятся детальные исследования параметров и гиперпараметров исследуемых моделей. Подробно описываются средства языка Python, начиная со среды разработки, заканчивая реализованными автором функциями. Приводятся результаты экспериментов на Python. Обосновывается выбор наиболее оптимальной модели, ее существенных признаков (features), определяется точность моделей и рассчитываются средневзвешенные значений параметров.

Глава 3. Разработка программных средств и проведение эксперимента на R. Сравнительная характеристика имплементации на Python и R. В данной главе приводятся результаты экспериментов,

реализованных посредством языка. Дается описанием использованных средств, функций и атрибутов для имплементации алгоритмов decision tree и random forest. Приводится сравнительная характеристика средств для имплементации на Python и R для выявления преимуществ и недостатков этих программных платформ.

Глава 4. Пользовательский интерфейс. Финальная глава показывает процесс создания веб приложения, на основе модели смешанного леса, полученной во второй главе. Подробно описывается создание как серверной (Python), так и клиентской частей приложения (JavaScript, HTML, CSS).

ЗАКЛЮЧЕНИЕ

В ходе данной работы были сделаны следующие выводы:

1. Деревья принятия решений являются отличным алгоритмом для целей принятия решений (decision making), в частности, для целей медицинской диагностики, а случайные леса – их эффективным дополнением, которое может при правильном подборе параметров значительно улучшить точность модели.

2. Критерии gini, entropy имеют различия во внутренней имплементации и интерпретации, несколько отличаются в плане производительности, однако не оказывают существенного влияния на обучение моделей.

3. Есть параметры, используемые для настройки моделей, которые оказывают большее влияние на обучение, а есть и те, которые не оказывают существенного влияния. К первым можно отнести, например, глубину дерева, ко вторым – минимальный размер выборки для разбития узлов.

4. Также устанавливается важность правильного разбиения набора данных, выбор объема тренировочных данных и требования к их качеству.

5. Показано, что в наборе данных нужно выделять признаки, которые имеют первоочередное значение (в нашей работе это уровень глюкозы, индекс массы тела и возраст), а также имеющие меньший приоритет.

6. Продемонстрировано, что Python является в целом более эффективным, нежели R для целей машинного обучения и последующей интеграции в веб-контент, хотя язык R также является весьма практически приемлемым выбором. Средства для разработки моделей деревьев в обоих языках имеют определённые отличия, свои плюсы и минусы.

7. Расчеты средневзвешенной оценки вероятности заболевания продемонстрировали мощность и предпочтительность случайных лесов. Кроме

того, анализ данных коэффициентов продемонстрировал, что точность модели не всегда находится в прямой зависимости от важности признака из набора данных.

8. Следует отметить, что в республике практически нет доступных веб ресурсов, где бы можно было пройти диагностику на диабет. Более того, неизвестно, созданы ли они на основе достижений науки о данных (data science) или построены на основе иных платформ классического программирования. В диссертации использован, пожалуй, лучший алгоритм в области машинного обучения для принятия решений и оценки вероятности наличия заболевания.

Результатом данной работы стала диагностическая система, построенная на основе детального анализа деревьев принятия решения и случайных лесов и последующего обучения моделей на основе данных алгоритмов, созданная для медицинских целей.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

По теме диссертации опубликована статья: Пилипенко В. Консультационная медицинская система / Н. В. Снатович, В. В. Пилипенко // Секция «Автоматизированные системы обработки информации»: программа конференции «Информационные технологии и управление 2021».