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

УДК 519.24;616-08

Корховая
Алина Богдановна

Алгоритмы анализа многомерных данных на основе медицинских показаний

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Программная инженерия

Научный руководитель
Абрамович М.С.
к.ф.-м.н., доцент

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

Рак лёгких является одним из наиболее распространённых видов онкологических заболеваний. Статистика по онкозаболеваниям в Беларуси, как и во всём мире с каждым годом растёт. В нашей стране ежегодно количество новых онкослучаев растёт примерно на тысячу человек.

Злокачественную опухоль лёгкого трудно вылечить, потому что сложно выявить на ранней стадии. Заболевание длительное время протекает бессимптомно. А для обнаружения и диагностирования рака легких часто требуется проведение нескольких исследований. На основании информации, полученной при проведении этих исследований, можно определить тип опухоли и ее распространение, а также можно планировать наилучшее возможное лечение для пациента.

Сейчас существует множество возможностей улучшить процесс лечения и диагностирования сложных болезней. С использованием информационных технологий в медицинской сфере намного выросли объемы обрабатываемой и хранимой информации. На данный момент самыми популярными медицинскими технологиями являются системы обработки снимков легких, головного мозга и другие. Однако есть множество необработанных статистических данных о пациентах с различными заболеваниями, в том числе рака лёгкого.

После диагностирования рака лёгких перед врачом предстает выбор пути лечения болезни. После консультации со специалистом было выявлено, что исход лечения, а, следовательно, и болезни, влияют множество факторов. Данным специалистом была предоставлена тестовая выборка, которая включала огромное количество признаков.

И так как медицинские показатели являются разнотипными, то для прогнозирования успеха лечения можно использовать следующий метод интеллектуального анализа данных: деревья решений или лес деревьев решений. Так же для первоначальной работы с данными были отобраны признаки при помощи специалиста, а затем необходимо применить алгоритм отбора информативных признаков.

Рак легкого – одна из основных онкологических проблем в Республике Беларусь. На протяжении последних лет в Республике Беларусь постепенно улучшаются показатели распределения вновь выявленных случаев рака по стадиям заболевания, и это происходит за счет увеличения частоты выявления случаев злокачественных новообразований в I – II стадиях (с 38,6 до 59,9 % к числу вновь выявленных случаев за указанный период) практически при всех основных локализациях.

Однако при раке легкого этот показатель увеличился незначительно – с 29,6 % в середине 80-х гг. до всего лишь 34,5 % в 2008 г., и увеличение произошло не за счет уменьшения процента больных, у которых заболевание выявлено в III и IV стадиях, а за счет уменьшения числа больных, стадия опухолевого процесса у которых не установлена. В структуре онкологической смертности мужчин рак легкого занимает первое место и в 2008 г. составил 27,0 %.

К настоящему моменту лечение рака легкого показывает свою эффективность и на территории Беларуси, однако своевременное лечение и диагностирование может привести к более благоприятным показателям. На практике было выявлено, что при любом типе онкологии ранее обнаружение болезни позволяет значительно улучшить выживаемость пациента.

Повысить шанс раннего прогноза заболевания можно, применяя современные диагностические технологии, а также по возможности используя информационные технологии и различные механизмы обработки данных.

Результаты данного исследования призваны помочь улучшить процесс выбора лечения для пациента с диагностированным заболеванием рака легких, опираясь на значения медицинских показателей.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Объектом исследования являютсяметоды машинного обучения в медицинской сфере. Исследования в данной сфере не теряют своей актуальности и по сей день. Каждый год публикуется огромное количество статей и исследований по различным направлениям. В данной работе исследуется сфера онкологии рака легких и возможность применения статистических и технологических решений для получения полезных результатов, которые могут улучшить лечение болезни.

Применение технологических решений в сфере здравоохранения с каждым годом растет и развивается. В рамках магистерской диссертации рассматриваются случаи из минской области, предоставленные работником сферы здравоохранения. Особенно актуально увеличение использования таких решений на территории Республики Беларусь.

Цель и задачи исследования

Цели диссертационной работы:

изучения алгоритмов машинного обучения для обработки многомерных данных;

реализация алгоритмов для оценки результата лечения рака лёгких;

исследование эффективности применения реализованных методов машинного обучения для оценки результата лечения рака лёгких.

Для достижения поставленной цели необходимо решить следующие задачи:

провести аналитический отбор применения методов машинного обучения для диагностики заболеваний;

осуществить программную реализацию алгоритмов машинного обучения на языке PYTHON версии не ниже 3.5;

определить параметры алгоритмов машинного обучения и ансамблей алгоритмов, при которых будет достигнута максимальная точность оценки результата лечения рака лёгких;

провести сравнительный анализ эффективности различных моделей классификаторов для оценки результата лечения рака лёгких.

Объектом исследования выступают методы машинного обучения для оценки результата лечения рака лёгких.

Предметом исследования является использование алгоритмов машинного обучения для анализа медицинских данных на примере данных о лечении рака лёгких.

Основной *гипотезой*, положенной в основу диссертационной работы, является получение аналитического результата обработки медицинских данных на примере данных о лечении рака легких.

Новизна полученных результатов

В результате проведенного исследования впервые были проанализированы статистические данные по пациентам с онкологией легких на территории Минской области и были опубликованы результаты исследования.

Также в результате работы были предложены алгоритмы классификации для определения успешности исхода лечения пациентов. Из предложенных алгоритмов был выбран алгоритм с самыми лучшими показателями классификации.

Положения, выносимые на защиту

1. Применение статистических методов обработки данных для анализа данных по пациентам с онкологией легких.
2. Применение алгоритмов классификации и их ансамблей для оценки применённого лечение и исходе заболевания пациента.

Апробация результатов диссертации

Материалы, положенные в основу работы, докладывались и обсуждались на седьмой международной научно-практической конференции «BIGDATAandAdvancedAnalyticsConferenceandEXPO» (Минск, Беларусь, 2021) и на 56-ой и 57-ой научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, Беларусь, 2020 и 2021).

Опубликованность результатов диссертации

По теме диссертации опубликовано 3 печатных работы в сборниках материалов международных научных конференций. Из них 1 работа в рецензируемом сборнике материалов VII-ой международной-практической конференции «BIGDATAandAdvancedAnalyticsConferenceandEXPO и 2 работы в сборниках трудов и материалов научных конференций БГУИР.

Структура и объем диссертации

Структура диссертации обусловлена целью, задачами и логикой исследования. Работа состоит из перечня условных обозначений и терминов, общей характеристики работы, введения, трёх глав, заключения, списка использованных источников, списка публикаций автора и приложения.Общий объем работы составляет 63страницы, из которых основного текста – 50 страниц, 16 рисунковна

14 страницах и7 таблиц на 7 страницах, список использованных источников из 26 наименований на 2 страницах и 1 приложение на 4 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе диссертации представлен обзор существующих алгоритмов анализа многомерных данных, используемых в медицинской сфере. В первом разделе произведен обзор информационных систем, которые используются в медицинской сфере для определения и исследования онкологий. Во втором разделе представлен обзор методов и подходов, которые используются в сфере здравоохранения. В конце главы представлены краткие выводы обзора и оценка ситуации в республике Беларусь.

Вторая глава включает в себя разработку подходов и выбора алгоритмов для анализа многомерных данных на основе медицинских показаний. В первом разделе представлены общие подходы к анализу многомерных данных. Во втором разделе рассмотрены основные понятия DataMining. В третьем – представлены возможные алгоритмы для анализа многомерных данных. В четвертом разделе детально рассматриваются алгоритмы Boosting, Backingи Stacking. В пятом разделе выполнен подбор алгоритмов для отбора информативных признаков. В кратких выводах второй главы финально были выбраны алгоритмы, которые будут применены в экспериментальной части работы.

Третья глава посвящена разработке алгоритмов анализа многомерных данных на основе медицинских показаний и разбор результатов. В первом разделе представлен анализ и преобразование полученной выборки данных для исследования. Во втором разделе описано применение алгоритмов машинного обучения для анализа многомерных данных на основе медицинских показаний и итоги их использования. В кратких выводах представлена оценка полученных результатов.

ЗАКЛЮЧЕНИЕ

В ходе работы над магистерской диссертацией была глубоко изучена проблематика исследования: важность использования математических и статистических методов анализа данных в медицине, а также использование современных информационных технологий для решения различных медицинских задач. Были проанализированы подходы в работе с большим объемом данных: подходы к их анализу, datamining, алгоритмы для анализа (регрессия, дерево решений, лес деревьев решений, Bagging, Boosting, Stacking).

Над предоставленными данными для работы была проведена работа по отбору параметров для анализа и разработки алгоритмов при помощи специалиста из сферы здравоохранения. Был выделен признак для предсказания и была поставлена задача определить успешность выздоровления пациента при определенном наборе параметров.

Для получения информативного набора этих параметров был применен критерий согласия Пирсона и были рассчитаны такие значения как X^2 и p-значение. Основываясь на втором значении и были отобраны информативные признаки для дальнейшего применения алгоритмов: проводилась ли лучевая терапия проводилась ли химиотерапия, какая операция проводилась по удалению раковой опухоли, были ли жалобы у пациента, гистологическая классификация рака лёгкого, стадия рака, диагностированная пациенту, без уточнения.

Так как задача была поставлена как задача классификации были применены алгоритмы классификации больного пациента на две группы: «успешный исход заболевания» и «неуспешный». Зависимым признаком был определен «Исход заболевания».

При анализе полученного датасета стало понятно, что выборка данных довольна мала (менее 400 пациентов). Исходя из этого для решения задачи классификации были выбраны алгоритмы дерево решений, случайный лес и ансамбль случайного леса и градиентного бустинга. После применения данных алгоритмов были проанализированы полученные результаты.

Наибольшая точность прогнозирования успешности лечения больных раком легкого была достигнута на экзаменационной выборке с применением смешанной модели и составила 90.5% на экзаменационной выборке. Этот же метод оказался более эффективным при классификации всей выборки больных раком легкого (89.2%).

Таким образом можно сделать вывод о целесообразности использования данных алгоритмов для решения поставленной задачи классификации пациентов с диагностированной онкологией легких, при этом ансамбли классификации показывают наиболее высокую точность предсказания.

Стоит отметить, что результаты работы на данный момент не имеют программной реализации для использования сторонними пользователями, однако в дальнейшем есть потенциал использовать полученные результаты исследования для помощи выбора методов решения для пациентов с диагнозом «рак легкого».

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. 56-я Научная Конференция Аспирантов, Магистрантов и Студентов – Корховая А.Б. – ИСПОЛЬЗОВАНИЕ АЛГОРИТМА ДЕРЕВА РЕШЕНИЙ ДЛЯ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ НА ПРИМЕРЕ ДАННЫХ ПО ОНКОЛОГИЧЕСКИМ ЗАБОЛЕВАНИЯМ ЛЁГКИХ. – 64-66 стр. – 21–24 апреля 2020 года, Минск, БГУИР [Электронный ресурс]. – Режим доступа: https://www.bsuir.by/m/12_100229_1_144999.pdf

2. 57-я Научная Конференция Аспирантов, Магистрантов и Студентов – Корховая А.Б. – АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ НА ПРИМЕРЕ ДАННЫХ ПО ОНКОЛОГИЧЕСКИМ ЗАБОЛЕВАНИЯМ ЛЁГКИХ: ПОДБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ. –19 –23 апреля 2021 год, Минск, БГУИР

3. Седьмая международная научно-практическая конференция BIG DATA andAdvancedAnalyticsConferenceand EXPO - А.Б. Корховая, М.С. Абрамович

Прогнозирование успешности лечения рака легкого с применением ансамблей классификаторов – 412-414 с. – 19-20 Мая, 2021 год Минск, Беларусь [Электронный ресурс]. – Режим доступа: http://bigdataminsk.bsuir.by/files/2021_stati.pdf