Использование методов машинного обучения для идентификации типа вируса

A Preprint

Анастасия Алексеевна Елисова
Факультет ВМК
МГУ имени М.В.Ломоносова
Москва, Россия
s02210381@gse.cs.msu.ru

Олег Валентинович Сенько
Факультет ВМК
МГУ имени М.В.Ломоносова
Москва, Россия
senkoov@mail.ru

Abstract

Данная статья посвящена выявлению зависимостей респираторных вирусных инфекций: гриппа, аденовируаса, респираторно-синцитиального вируса и бокавируса от предоставленных данных (симптомов) на основе ансамблевых методов, а также тестов. Основным методом решения задач компьютерной диагностики является обучение алгоритма машинного обучения на основе выборки, которая состоит из описания отдельных пациентов. Однако такая информация часто ограничена рамками одного медицинского учреждения и имеет небольшой объем. В результате настоящей работы были проведены эксперименты и вычисления, позволяющие выявить статистическую и линейную зависимости между данными.

Keywords Машинное обучение · Статистический анализ · Вирусы

1 Introduction

Острые респираторные вирусные инфекции (ОРВИ) оказывают значительное влияние на системы здравоохранения по всему миру из-за высокой заразности и уровня госпитализаций, особенно среди детей, пожилых и людей с ослабленным иммунитетом. Эти инфекции являются одной из самых частых причин обращения за медицинской помощью, занимая лидирующие позиции по заболеваемости и смертности среди инфекционных заболеваний. Диагностика и дифференцировка этих инфекций по симптомам затруднены, особенно в условиях ограниченности данных, что ставит перед медицинскими специалистами задачу правильного и своевременного распознавания типа вируса. Ансамблевые методы машинного обучения и статистический анализ помогают повысить точность диагностики, а также выявить статистически значимые взаимосвязи между симптомами и вирусами.

Современная диагностика респираторных инфекций в основном базируется на клинических данных, таких как жалобы пациентов, анамнез, физикальные исследования, а также лабораторных тестах. Однако в условиях массовых обращений и недостатка лабораторных ресурсов точность диагностики может быть низкой. Традиционные методы, такие как ПЦР-диагностика или серологические тесты, зачастую требуют времени и значительных финансовых затрат. С этим связана необходимость внедрения более быстрых и доступных методов диагностики, например, с использованием алгоритмов машинного обучения, которые могут анализировать большое количество данных и предоставлять результаты с высокой точностью и в краткие сроки.

Вирус гриппа типа А — один из наиболее опасных вирусов, вызывающих как сезонные эпидемии, так и пандемии. Он имеет высокую изменчивость антигенов, что делает невозможным долгосрочное создание универсальной вакцины. Основные симптомы включают высокую температуру, озноб, слабость, боль в мышцах и головную боль, что отличает его от большинства других ОРВИ. Из-за высокой мутабельности

вируса, эпидемии происходят с периодической сменой штаммов, что требует регулярного обновления вакцин.

Аденовирус — вирус, вызывающий широкий спектр инфекций, включая респираторные, глазные и кишечные заболевания. Симптомы включают лихорадку, боль в горле, кашель, насморк и конъюнктивит. Аденовирусы могут поражать как верхние, так и нижние дыхательные пути и устойчивы к неблагоприятным условиям окружающей среды, что способствует вспышкам в закрытых детских учреждениях и медицинских организациях. Эти вирусы могут вызывать тяжёлые формы заболеваний у людей с ослабленным иммунитетом, таких как пациенты с онкологическими заболеваниями или трансплантированными органами.

Бокавирус — относительно недавно обнаруженный вирус, часто связанный с острыми респираторными инфекциями, особенно у детей. Он может вызывать симптомы, напоминающие грипп и РСВ, такие как кашель, лихорадка и затруднённое дыхание. Бокавирус часто сосуществует с другими вирусами, что затрудняет его диагностику и требует комплексного подхода для точного распознавания. Особенно важно учитывать его наличие при лечении пневмоний и бронхитов, когда обычные методы диагностики могут не дать точных результатов.

Риновирус — основная причина простуды и одна из наиболее частых причин заболеваний верхних дыхательных путей. Симптомы риновирусной инфекции включают насморк, боль в горле, головную боль и иногда кашель. В отличие от других ОРВИ, риновирусы редко сопровождаются высокой температурой, что делает их диагностику затруднённой, особенно в зимний период, когда происходит повышенная циркуляция других вирусов.

Респираторно-синцитиальный вирус (РСВ) — вирус, вызывающий инфекции нижних дыхательных путей, особенно у младенцев и пожилых людей. Он является одной из ведущих причин бронхиолита и пневмонии. Основные симптомы РСВ включают лихорадку, кашель, хрипы и затруднённое дыхание, что требует своевременной диагностики для назначения адекватного лечения. Существует также повышенный риск осложнений у людей с хроническими заболеваниями лёгких и сердечно-сосудистой системы.

Настоящее исследование сосредоточено на использовании ансамблевых методов машинного обучения для повышения точности диагностики респираторных вирусных инфекций. Целью работы является выявление ключевых симптомов, отличающих каждую инфекцию, и разработка модели, которая на основе этих симптомов будет эффективно распознавать тип вируса. В ходе исследования используются статистические методы для оценки значимости симптомов, а также алгоритмы машинного обучения для построения классификатора, который поможет улучшить диагностику и лечение ОРВИ в клинической практике. Модели машинного обучения, такие как случайные леса, градиентный бустинг и нейронные сети, способны эффективно обрабатывать большое количество признаков и выявлять сложные зависимости между симптомами и вирусами. На основе этих данных могут быть построены системы раннего оповещения и диагностики, что поможет снизить заболеваемость и смертность от ОРВИ.

2 Постановка задачи

В данной работе использована таблица с информацией о характеристиках гриппа, аденовируса, респираторно-синцитиального вируса и бокавируса у 1533 пациентов. Симптомы, рассматриваемые в данном исследовании, включают температуру, кашель, одышку, боли в горле, боль в груди, ринит, головную боль, диарею боли в животе, слабость и многие другие.

Для оценки качества используется метрика ROC-AUC, как рекомендуемая Министерством Здравоохранения для использования в доказательной медицине.

3 Анализ вирусов

	Количество нулей	Количество единиц
Адено	108	1425
Бока	62	1471
Рино	313	1220
ГриппА	157	1376
РС-вирус	211	1322

Таблица 1: Соотношение количества позитивных и негативных ответов для всех типов вирусов

Таким образом, можно сделать вывод о том, что данные не сбалансированы.

Тип вируса / Симптом	Грипп	РС	Адено	Бока	Рино
Температура					
Кашель					
Одышка					
Боли в горле					
Боль в груди					
Ринит					
Головная боль					
Диарея					
Боли в животе					

Таблица 2: Информация о количестве зафиксированных симптомов для каждого из заболеваний

4 Анализ симптомов

4.1 Точный тест Фишера

Точный тест Фишера используется для определения того, существует ли значимая связь между двумя категориальными переменными. Обычно он применяется в качестве альтернативы критерию независимости Хи-квадрат.
Тест независимости по методу Хи-квадрат - это более традиционная проверка гипотезы, которая использует тестовую статистику (хи-квадрат) и ее выборочное распределение для вычисления p-значения. Однако, распределение выборки по методу хи-квадрат только приближает правильное распределение, обеспечивая лучшие значения p по мере увеличения значений ячеек в таблице. Следовательно, значения p по хи-квадрату недопустимы при малом количестве ячеек.
Точный тест Фишера использует следующие нулевые и альтернативные гипотезы:

- H_0 (нулевая гипотеза): Две переменные независимы.
- H_1: (альтернативная гипотеза): Две переменные не независимы.

По имеющимся значениям строится следующая табличка:

Однозначное значение p для точного критерия Фишера рассчитывается как:

$$p\_value = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Если значение p ниже установленного уровня значимости (в данной работе - 0,05), нулевая гипотеза отклоняется. Данные выборки достаточно убедительны, чтобы сделать вывод о существовании взаимосвязи между категориальными переменными в генеральной совокупности.

Ниже представлены результаты подсчёта пи-значениения для бинарных признаков в настоящей задаче (зеленым цветом выделены существенные значения).

	Грипп	РС	Грипп	Адено	Грипп	Бока	РС	Адено	РС	Бока	Адено	Бока

Таблица 3: Точный критерий Фишера, полученный по таблице сопряженности для каждых двух вирусов и каждых симптомов

4.2 Тест Манна-Уитни

4.3 Перестановочный тест

Список литературы

[1] Материалы ММКЦ "Коммунарка

[2] Изучение влияния клинико-генетических факторов на течение дисциркуляторной энцефалопатии с использованием методов распознавания: Кузнецова А.В., Костомарова И.В., Водолагина Н.Н., Малыгина Н.А., Сенько О.В.

[3] Сологуб Т.В., Осиновец О.Ю. Иммуномодуляторы в ком- плексной терапии ОРВИ: возможности применения препа- рата галавит. Русский медицинский журнал. 2013;3.

[4] Инфекционные болезни: национальное руководство. Под общ. ред. Ющука Н.Д., Венгерова Ю.Я. М.: Гэотар-Мед.; 2009.

[5] Empirical testing of institutional matrices theory by data mining I. L. Kirilyuk, A. I. Volynsky, M. S. Kruglova, A. V. Kuznetsova, A. A. Rubinstein, O. V. Senko.