Использование методов машинного обучения для идентификации типа вируса

В этой презентации мы рассмотрим, как методы машинного обучения могут помочь в точной идентификации различных типов вирусов, таких как грипп, аденовирус, респираторно-синцитиальный вирус и бокавирус. Мы изучим эффективность статистических тестов, таких как тест Фишера, тест Манна-Уитни и перестановочный тест, а также метод статистически взвешенных синдромов в диагностике вирусных инфекций.

Елисова Анастасия Алексеевна

15 ноября 2024г.

Актуальность проблемы

1

Быстрая диагностика

Своевременная и точная диагностика вирусных инфекций имеет решающее значение для предотвращения распространения заболеваний и назначения подходящего лечения.

2

Эпидемиологический контроль

Идентификация типов вирусов помогает отслеживать тенденции распространения и вырабатывать эффективные меры общественного здравоохранения.

3

Разработка вакцин

Понимание вирусных штаммов критически важно для создания таргетированных вакцин и укрепления общественного иммунитета.

Обзор существующих методов диагностики

Традиционные методы

Методы на основе культивирования вирусов и иммунологические тесты, такие как ИФА и РИФ, широко используются, но имеют ограничения по времени и точности.

Молекулярно-генетические методы

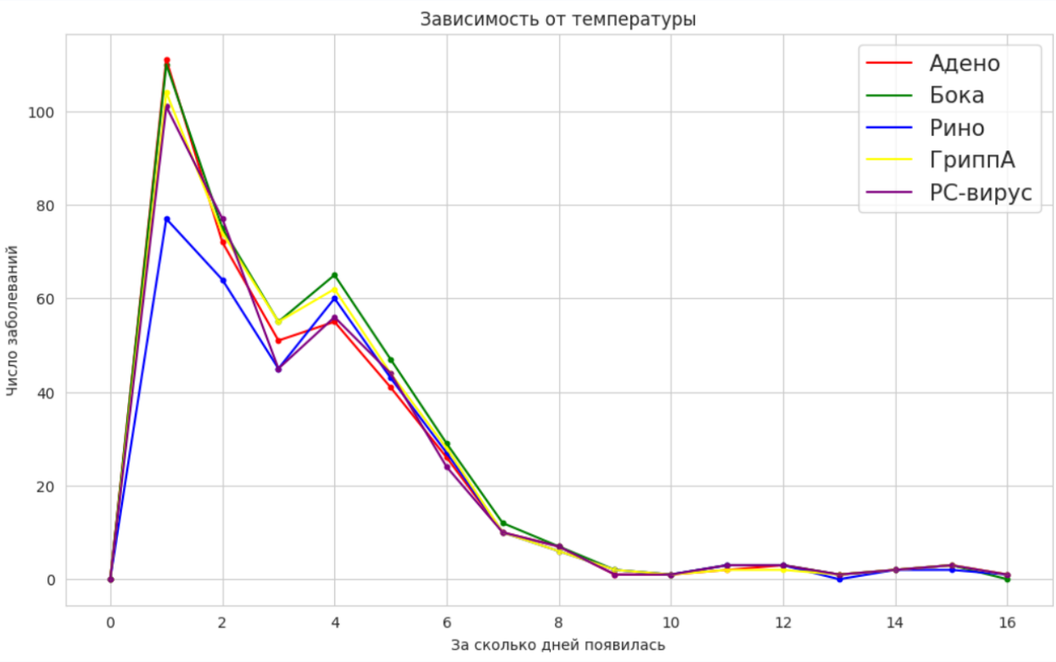
Полимеразная цепная реакция (ПЦР) и секвенирование обеспечивают высокую точность, но требуют сложного оборудования и специальных навыков.

Методы машинного обучения

Новые алгоритмы анализа биомаркеров и симптомов позволяют быстро и точно идентифицировать различные типы вирусов.

Описание набора данных

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



Источник данных

Набор данных был получен из ММКЦ "Коммунарка" и представляет результаты обследования пациентов с различными вирусными инфекциями.

Объем данных

Данные включают в себя информацию о 1533 пациентах, у которых были выявлены четыре основных типа вирусов: гриппА, риновирус, аденовирус, респираторно-синцитиальный вирус и бокавирус.

Переменные

Набор данных содержит информацию о симптомах, демографических характеристиках, результатах лабораторных анализов и исходе заболевания для каждого пациента.

Применение статистических тестов

Тест Фишера

Этот точный тест используется для анализа различий в частотах категориальных переменных между группами пациентов с разными вирусными инфекциями.

Тест Манна-Уитни

Непараметрический тест, который позволяет сравнивать распределения непрерывных переменных, таких как лабораторные показатели, между группами.

Перестановочный тест

Метод, основанный на рандомизации данных, дает возможность оценить статистическую значимость различий между группами без строгих предположений о распределении.

Тест Фишера

Точный тест Фишера использует следующие нулевые и альтернативные гипотезы:

- H0 (нулевая гипотеза): Две переменные независимы.
- H1 (альтернативная гипотеза): Две переменные не независимы.

По имеющимся значениям строится следующая табличка:

	Group 1	Group 2	Row Total
Category 1	a	b	a+b
Category 2	c	d	c+d
Column Total	a+c	b+d	a+b+c+d = n

Где a - количество случаев, когда обе переменные принимают определенные категории (например, да/да), b - количество случаев, когда первая переменная принимает одну категорию, а вторая — другую, c - количество случаев, когда первая переменная принимает другую категорию, а вторая — первую, d - количество случаев, когда обе переменные принимают противоположные категории.

$$p_{value} = \frac{(a\ +\ b)!\,(c\ +\ d)!\,(a\ +\ c)!\,(b\ +\ d)!}{a!\,b!\,c!\,d!\,n!}$$

	ЧБД	Аллергоанамнез	Антибиотикинадогоспэтапе	Температура37-37,9	Температура38-39	температуравыше39,1	Слабость	головнаяболь	Мышечнаяболь	кашель
Адено	0.694798	0.034326	0.106807	0.058522	0.724977	0.000160	0.250843	0.375466	1.000000	0.000000
Бока	0.303352	0.541860	0.765073	0.078072	0.541186	0.378364	0.352762	0.257810	1.000000	0.000066
Рино	0.559318	0.000341	0.341007	0.053765	0.004546	0.008819	1.000000	0.099648	1.000000	0.000053
ГриппА	0.375182	0.028982	0.079896	0.859641	0.012971	0.908224	0.010300	0.809586	0.013584	0.112970
РС-вирус	0.170856	0.001241	0.347133	0.119859	0.793853	0.041301	0.001268	0.001787	0.374557	0.000000

Тест Манна-Уитни

Тест Манна-Уитни используется для проверки гипотезы о том, что две выборки взяты из распределений с одинаковыми медианами, или что одно распределение доминирует над другим, то есть, значения из одной группы чаще больше значений из другой группы. Это особенно полезно, когда данные не соответствуют предположению о нормальности, и, следовательно, не могут быть проанализированы с помощью стандартного t-теста.

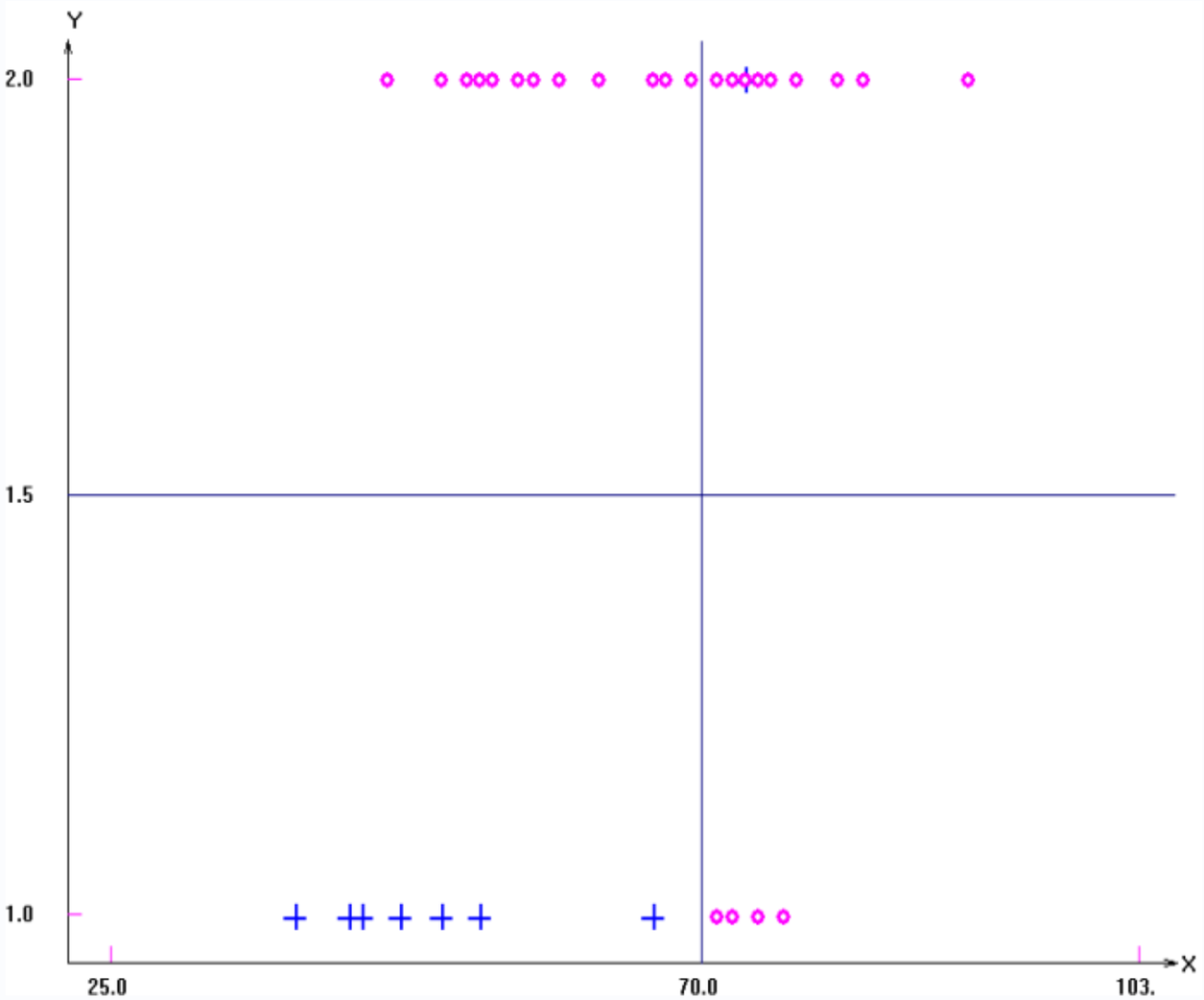
Шаги выполнения теста Манна-Уитни:

1. Объединяем все наблюдения из обеих выборок и ранжируем их. Т.е. для каждого элемента вычисляется его ранг в общем наборе.
2. Для каждого наблюдения из группы 1 (и группы 2) присваиваем его ранг в объединенной выборке.
3. Вычисляем сумму рангов для каждой из групп. Обозначим сумму рангов для группы 1 как R1, а для группы 2 — R2.
4.Считаем статистику U для группы 1 (и для группы 2). Формулы для статистики U выглядят следующим образом:

$$U_1 = R_1 - \frac{m(m+1)}{2} \qquad U_2 = R_2 - \frac{n(n+1)}{2} \qquad \mathrm{U} = \min(U_1, U_2)$$

	возраст год	Деньболезни	Температураколькоднейдопоступленни	Эози нофилы %	П/ядерные %	моноциты %	Эозинофилы	моноциты	Мочевина.1	Креатинин.1
Адено	0.000003	0.000025	0.011808	0.000029	0.000163	0.000075	0.000035	0.000102	0.000027	0.000021
Бока	0.000001	0.000003	0.000225	0.000533	0.000012	0.000039	0.000009	0.000061	0.000000	0.000000
Рино	0.002273	0.046081	0.028801	0.049020	0.015154	0.002841	0.071484	0.015527	0.000000	0.000006
ГриппА	0.000023	0.000039	0.005980	0.000341	0.001000	0.003311	0.000001	0.000850	0.000000	0.000077
РС-вирус	0.000093	0.001092	0.012898	0.000647	0.000225	0.003355	0.000131	0.000559	0.000001	0.000006

Метод статистически взвешенных синдромов



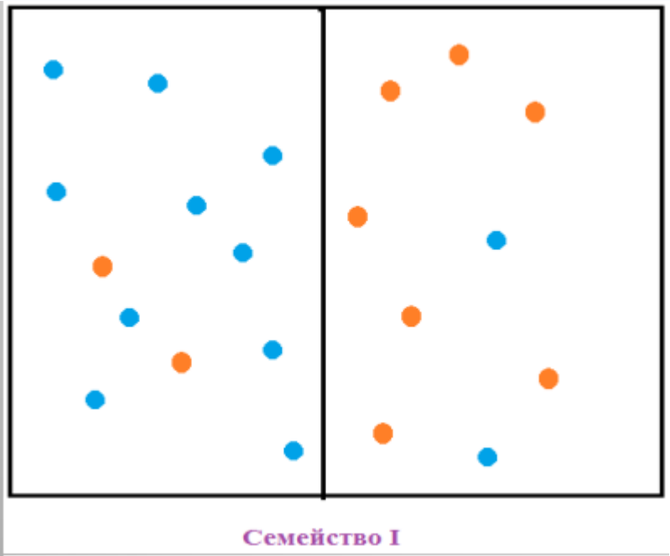
$$F_s(K_j, S_t, R) = \sum_{i=1}^{t} \left(v_0^j - v_i^j\right)^2 m_i$$

$$F_L(K_j, S_t, R) = \max_{i \in \{1..,t\}} \left(v_0^j - v_i^j\right)^2 m_i$$

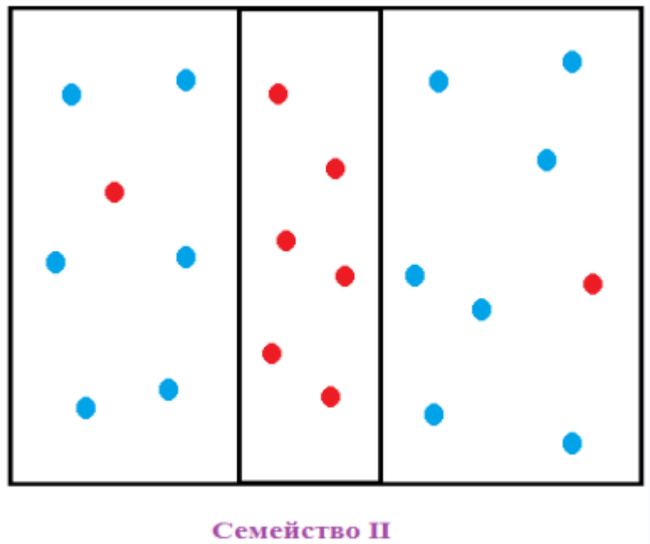
$$\Gamma_j(s) = \frac{\sum_{i=1}^{r} w_i v_i^j}{\sum_{i=1}^{r} w_i}$$

$$w_i = \frac{m_i}{m_i + 1} \frac{1}{v_i^j (1 - v_i^j)}$$

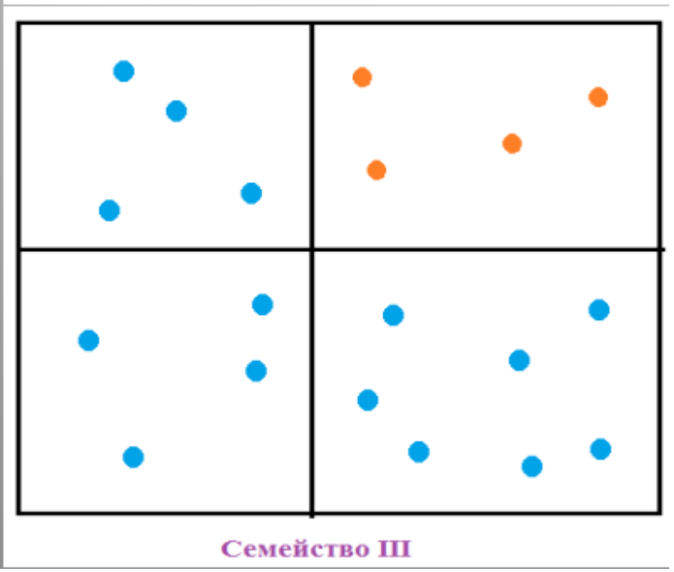
Метод статистически взвешенных синдромов



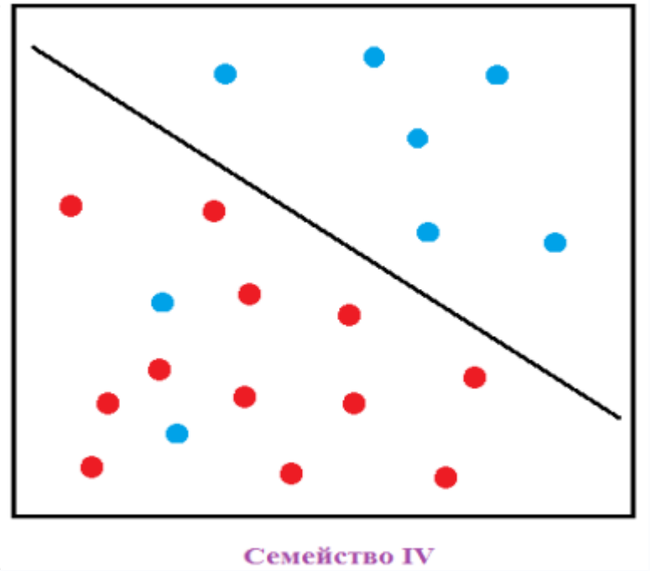
Семейство I



Семейство II



Семейство III



Семейство IV

Поиск разбиений с максимальным значением одного из функционалов производится в рамках одного из четырёх семейств. Примеры разбиений для каждого из семейств приведены на рисунке.

- *Семейство I* включает всевозможные разбиения интервалов допустимых значений отдельных признаков на два интервала с помощью одной граничной точки.
- *Семейство II* включает всевозможные разбиения интервалов допустимых значений отдельных признаков на 3 интервала с помощью двух граничных точек.
- *Семейство III* включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 4 подобласти с помощью двух граничных точек (по одной точке для каждого из двух признаков)
- *Семейство IV* включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 2 подобласти с помощью прямой граничной линии, произвольно ориентированной относительно координатных осей.

Практическое применение и перспективы

1

Быстрая диагностика

Предложенные методы позволяют в кратчайшие сроки идентифицировать тип вирусной инфекции, что критически важно для своевременного назначения лечения.

2

Повышение доступности

Автоматизация процесса диагностики снижает потребность в специализированном оборудовании и экспертных знаниях, расширяя доступ к качественной медицинской помощи.

3

Эпидемический контроль

Точная идентификация вирусных штаммов способствует лучшему эпидемиологическому надзору и выработке эффективных мер общественного здравоохранения.

Представленные подходы на основе машинного обучения открывают новые возможности для повышения качества и доступности диагностики вирусных инфекций. Дальнейшее развитие этих методов, их внедрение в реальную клиническую практику и интеграция с другими передовыми технологиями имеют большой потенциал для улучшения системы здравоохранения.