Мета-анализ и метод классификации в условиях отсутствия табличной обучающей выборки

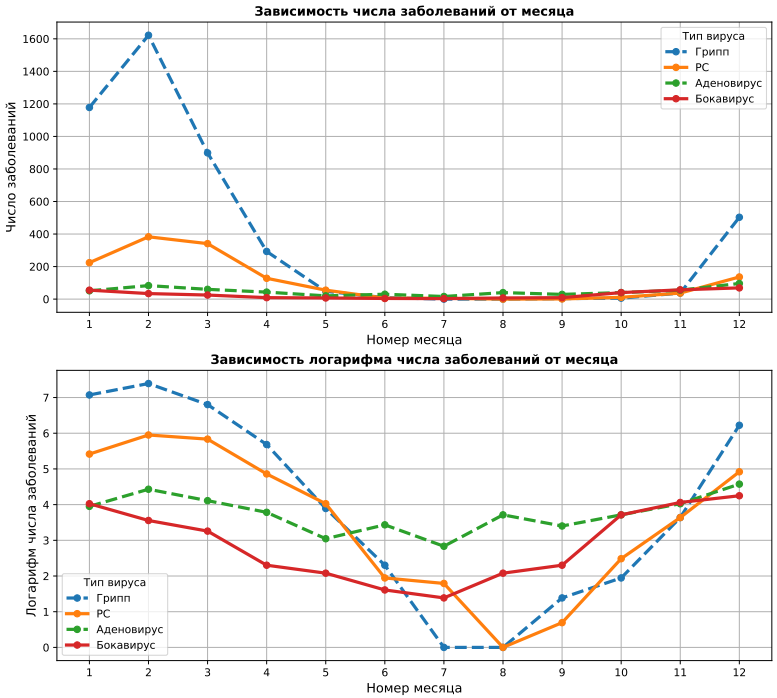
Полина Юрьевна Кривуля
Олег Валентинович Сенько
Воронин Евгений Михайлович
Плоскирева Антонина Александровна
Хлыповка Юлия Николаевна

2023

Введение

Основным методом решения задач компьютерной диагностики является обучение алгоритма машинного обучения на основе выборки, которая состоит из описания отдельных пациентов.
Проблема: информация часто ограничена рамками одного медицинского учреждения и имеет небольшой объем.
Цель работы: анализ данных, полученных из литературных источников и диагностика вирусных инфекций по набору признаков. Задача состоит в построении алгоритма, распознающего один из типов вирусов на основании набора симптомов и сезона.

Описание данных



Для получения априорной вероятности наличия вируса в определенный сезон использована отдельная таблица, отражающая заболеваемость каждым из вирусов по датам в течение нескольких лет.

Рассматривается база данных, собранных из медицинских статей, с характеристиками четырех вирусных инфекций. Для каждой статьи имеется информация о суммарном числе заболевших и о количестве людей, имеющих один из девяти симптомов.

Статистическая значимость признаков

Проводится анализ наличия зависимости между типом вируса и симптомом с помощью таблиц сопряженности и точного критерия Фишера. Установлен критический уровень значимости $p = 0.001$, все результаты, являющиеся статистически значимыми на этом уровне, выделены полужирным шрифтом.

	Грипп | РС	Грипп | Адено	Грипп | Бока	РС | Адено	РС | Бока	Адено | Бока
Температура	$\mathbf{3.7 * 10^{-4}}$	$\mathbf{1.4 * 10^{-33}}$	$\mathbf{7.8 * 10^{-25}}$	$\mathbf{3.5 * 10^{-31}}$	$\mathbf{8.2 * 10^{-10}}$	$\mathbf{6.8 * 10^{-62}}$
Кашель	$\mathbf{1.6 * 10^{-33}}$	$\mathbf{6.7 * 10^{-7}}$	$\mathbf{6.0 * 10^{-21}}$	$\mathbf{1.4 * 10^{-13}}$	$9.9 * 10^{-2}$	$\mathbf{1.7 * 10^{-11}}$
Одышка	$\mathbf{2.8 * 10^{-88}}$	$1.6 * 10^{-1}$	$\mathbf{9.1 * 10^{-30}}$	$\mathbf{4.5 * 10^{-48}}$	$1.4 * 10^{-1}$	$\mathbf{5.1 * 10^{-24}}$
Боли в горле	$5.4 * 10^{-2}$	**0.0**	$7.1 * 10^{-1}$	$\mathbf{1.3 * 10^{-132}}$	1.0	$\mathbf{1.6 * 10^{-9}}$
Боли в груди	$4.3 * 10^{-1}$	–	–	–	–	–
Ринит	$\mathbf{1.2 * 10^{-135}}$	$\mathbf{6.3 * 10^{-137}}$	$\mathbf{3.2 * 10^{-38}}$	$9.9 * 10^{-2}$	$\mathbf{8.4 * 10^{-5}}$	$\mathbf{2.2 * 10^{-6}}$
Головная боль	$\mathbf{6.9 * 10^{-7}}$	$\mathbf{1.3 * 10^{-59}}$	–	$\mathbf{1.4 * 10^{-45}}$	–	–
Диарея	$\mathbf{1.7 * 10^{-12}}$	$\mathbf{2.5 * 10^{-45}}$	$\mathbf{2.6 * 10^{-34}}$	$6.1 * 10^{-1}$	$\mathbf{3.6 * 10^{-6}}$	$\mathbf{1.2 * 10^{-7}}$
Боли в животе	$1.5 * 10^{-2}$	$\mathbf{7.9 * 10^{-37}}$	$\mathbf{2.2 * 10^{-64}}$	$4.3 * 10^{-3}$	$\mathbf{6.5 * 10^{-30}}$	$\mathbf{2.2 * 10^{-31}}$
Зима	$\mathbf{9.3 * 10^{-27}}$	$\mathbf{1.6 * 10^{-46}}$	$\mathbf{7.1 * 10^{-16}}$	$\mathbf{1.9 * 10^{-9}}$	$3.9 * 10^{-2}$	$1.4 * 10^{-2}$
Весна	$\mathbf{6.7 * 10^{-18}}$	$9.7 * 10^{-3}$	$\mathbf{4.6 * 10^{-9}}$	$\mathbf{5.7 * 10^{-14}}$	$\mathbf{1.2 * 10^{-21}}$	$\mathbf{8.5 * 10^{-4}}$
Лето	$1.5 * 10^{-3}$	$\mathbf{2.0 * 10^{-76}}$	$\mathbf{8.5 * 10^{-12}}$	$\mathbf{4.6 * 10^{-37}}$	$\mathbf{4.5 * 10^{-5}}$	$\mathbf{7.2 * 10^{-8}}$
Осень	$\mathbf{3.8 * 10^{-10}}$	$\mathbf{1.2 * 10^{-85}}$	$\mathbf{2.6 * 10^{-96}}$	$\mathbf{4.9 * 10^{-33}}$	$\mathbf{1.1 * 10^{-45}}$	$\mathbf{3.3 * 10^{-4}}$

Гетерогенность включаемых исследований

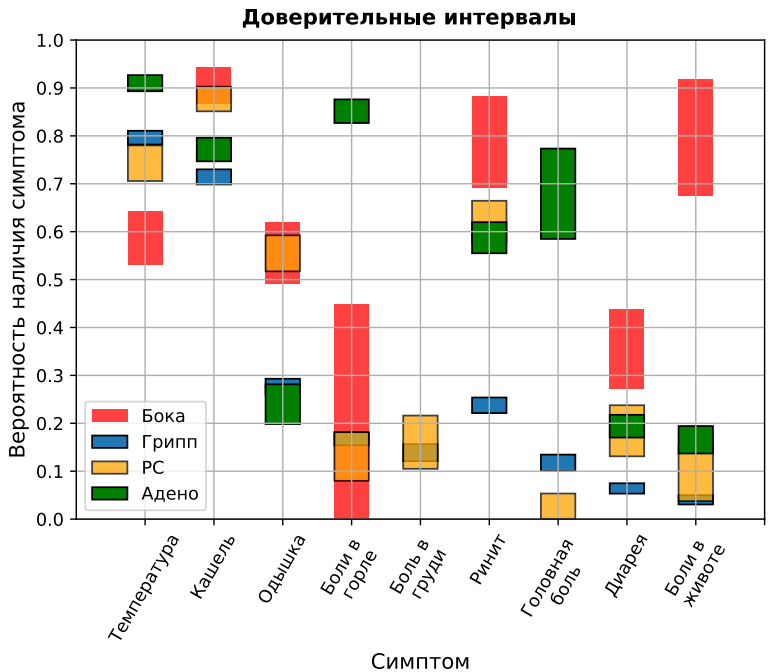
Рассматривается значение I^2, как наиболее широко применяемый показатель оценки гетерогенности исследований.

$$I^2 = \frac{Q - (N - 1)}{Q} * 100\%,$$

где N – число статей для рассматриваемого симптома и типа вируса, Q - значение статистики.
Пороговые значения низкой, умеренной и высокой гетерогенности исследований эмпирически установлены на уровнях 25, 50 и 75% соответственно.
Анализ показал, что данные, соответствующие одному типу инфекции, являются в значительной степени гетерогенными.

Гетерогенность включаемых исследований

Визуализированы доверительные интервалы Клоппера-Пирсона для бинарных ответов.



Предположение: гетерогенность не помешает работе алгоритма.

Алгоритм

Используем формулу Байеса для наивного байесовского классификатора:

$$P(K_l|x) = \frac{P(K_l)\prod_{i=1}^{n} p_l(x_i)}{\sum_{j=1}^{L} P(K_j)\prod_{i=1}^{n} p_j(x_i)} \qquad (1),$$

где K — тип вируса, n — число рассматриваемых симптомов, $x \in \{0,1\}^n$ — набор симптомов, $P(K_l)$ — априорная вероятность наличия вируса K_l в текущем сезоне, $p_j(x_i)$ — вероятность наличия (или отсутствия) симптома x_i у вируса j, L — число рассматриваемых вирусов.
У одного из вирусов отсутствует информация о наличии одного из рассматриваемых симптомов. Обозначим этот признак x_i.
Вирусы разделяются на две группы: G_1, состоящей из вирусов K_1, K_2, K_3, для которых известны все признаки, и G_2, состоящей из вируса K_4.

Алгоритм

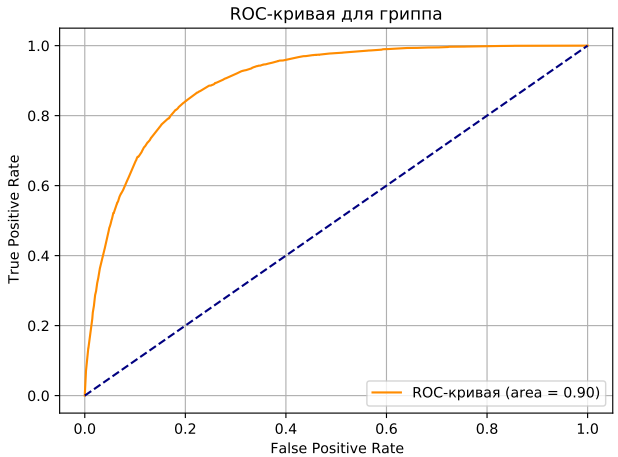
Шаг 1. Вычисление $P(G_1)$ и $P(G_2)$. Рассматривается подмножество признаков $\tilde{X}$, известных для всех вирусов. По этим признакам с использованием формулы (1) вычисляются вероятности наличия каждого из вирусов. Вероятность $P(G_1)$ вычисляется как сумма полученных вероятностей вирусов K_1, K_2, K_3; вероятность $P(G_2)$ как полученная вероятность вируса K_4.

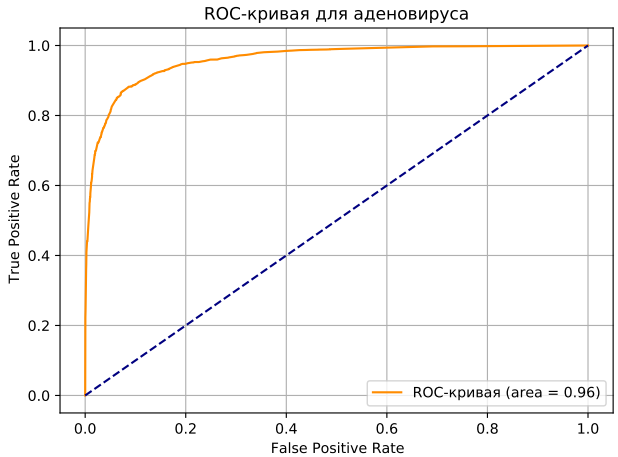
Шаг 2. Вычисление $P(K_1)$, $P(K_2)$, $P(K_3)$. Рассматриваются вирусы, для которых известны признаки из подмножества $\tilde{X}$ и признак x_i. С помощью формулы (1) с использованием всех признаков вычисляются вероятности $P(K_1)$, $P(K_2)$, $P(K_3)$. Полученные вероятности домножаются на значение $P(G_1)$, рассчитанное на шаге 1.

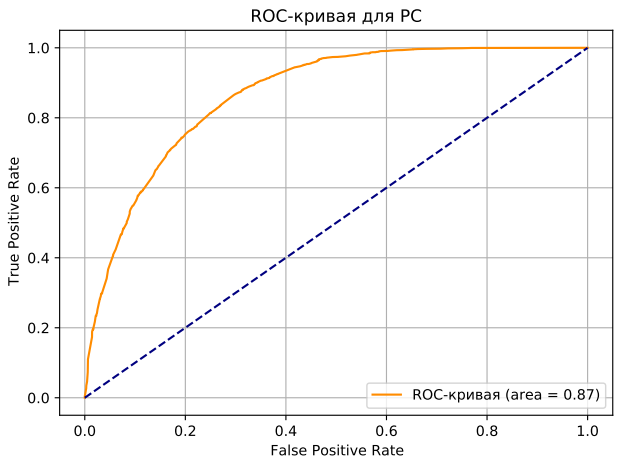
Шаг 3. Значение $P(K_4)$. Для вируса K_4 с отсутствующим симптомом x_i остается вероятность $P(G_2)$, вычисленная на шаге 1 с использованием формулы (1): $P(K_4) = P(G_2)$.

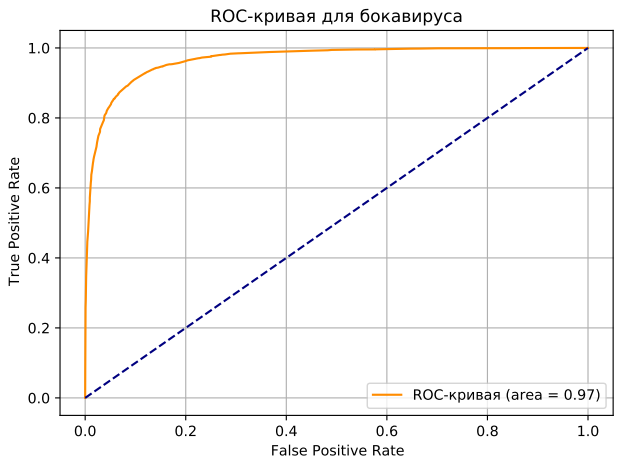
Результаты

Качество оценивается с помощью сгенерированной выборки и анализа ROC-кривой.









Заключение

Результаты

- Проведен анализ данных, собранных из медицинских статей;
- Построен алгоритм, позволяющий решать задачу классификации типа вируса по таким данным;
- Преимуществом метода является его актуальность в случаях, когда наличие полноценной обучающей выборки невозможно.