АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ НА ПРИМЕРЕ ДАННЫХ ПО ОНКОЛОГИЧЕСКИМ ЗАБОЛЕВАНИЯМ ЛЁГКИХ: ПОДБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Корховая А.Б.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Абрамович М.С. – зав. НИЛ статистического анализа и моделирования

В данной работе был рассмотрен и применен критерий согласия Пирсона χ2 для отбора информативных признаков выборки данных. Также был произведен выбор алгоритма для дальнейшей работы в рамках поставленной задачи.

По результатам наблюдений за онкологическими больными формируются выборки данных содержащие большое количество не всегда информативных признаков. Для проведения эксперимента была предоставлена выборка данных по Минскому району, которая включает в себя 380 пациентов и более 100 характеристик. Однако для такого объёма данных количество признаков несоизмеримо высоко. На первом этапе работы с данными было принято решение не использовать показатели генов. И далее при помощи эксперта были отобраны 11 параметров для которых будут производиться применения алгоритмов [1]. Следующим шагом в моей работе стал отбор информативных признаков при помощи математических алгоритмов.

Для отбора информативных признаков был выбран критерий согласия Пирсона χ2. Ведь это один из самых популярных статистических критериев для анализа качественных данных (номинальных, порядковых, ранговых), анализа частот.

Критерий согласия Пирсона применяют для проверки гипотезы о соответствии эмпирического распределения предполагаемому теоретическому распределению F(x) при большом объеме выборки (количество записей в выборке ≥ 100). Критерий применим для любых видов функции F(x), даже при неизвестных значениях их параметров, что обычно имеет место при анализе результатов механических испытаний. В этом заключается его универсальность.

Также данный алгоритм используется для [2]:

- для оценки статистической значимости различий двух или нескольких относительных показателей;
- для проверки гипотезы (нулевой), что наблюдаемая случайная величина подчиняется некому теоретическому закону распределения;
- для анализа категориальных данных, т.е. таких, которые выражаются не количеством, а принадлежностью к какой-то категории. К таким данным нельзя применять математические операции вроде сложения и умножения, для них можно только подсчитать частоты.

В таблице 1 приведены значения статистики хи-квадрат и соответствующие *p*-значения признаков. В качестве информативных признаков отбирались те, у которых *p*-значения меньше 0.1.

Таблица 1 – Информативные признаки

Признак	Статистика хи-квадрат	*p*-значение
Лучевая терапия	41.90860	0.000000
Химио-терапия	25.75409	0.000000
Операция (да; нет)	25.65689	0.000000
Жалобы (шифр)	12.82626	0.000342
Гист. РЛ (шифр)	18.90845	0.001999
Статус курения	7.16375	0.007439
Возраст	64.35080	0.024284
Стадия	8.86487	0.031142

Признаки упорядочены по p-значению. Чем меньше p-значение, тем признак более информативен.

Таким образом из одиннадцати признаков, выбранных экспертом, для дальнейшей работы были отобраны 8 информативных признаков. Которые в дальнейшем будут использоваться в алгоритме дерева решений. Деревья решений в свою очередь рекомендуется использовать в условиях выборок ограниченного объема [3], что и имеет место в рассматриваемом случае.

Список использованных источников:

1. Корховая А.Б. ИСПОЛЬЗОВАНИЕ АЛГОРИТМА ДЕРЕВА РЕШЕНИЙ ДЛЯ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ НА ПРИМЕРЕ ДАННЫХ ПО ОНКОЛОГИЧЕСКИМ ЗАБОЛЕВАНИЯМ ЛЁГКИХ. — 21–24 апреля 2020 года, Минск, БГУИР.

2. Никулин М. С. О критерии хи-квадрат для непрерывных распределений. — 1973

3. Harrington, P. Machine Learning in Action / P. Harrington. New York: Manning, 2012