Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

Национальный Исследовательский Университет

«Высшая Школа Экономики»

Факультет компьютерных наук

Магистерская программа Науки о Данных

Кафедра Анализа Интернет-Данных

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

На тему “Прогнозирование успеха таргетной терапии онкологических заболеваний при помощи данных клеточных линий"

Студент группы № мНоД16 АИД
Артемьев Михаил Александрович

Руководитель ВКР
к.ф.-м.н., доцент Федотов Станислав Николаевич

Москва, 2018

Содержание

1 Введение

Эффективность терапии онкологических заболеваний можно прогнозировать. Построение такого прогноза может помочь более точному назначению препаратов и, как следствие, ускорить лечение, минимизировать осложнения, связанные с побочными действиями препаратов, снизить стоимость лечения.

Для прогнозирования результата терапии онкологических заболеваний, в частности, используются активации сигнальных путей [1], [8] – признаки, вычисляемые по профилю генной экспрессии пациента. Объём выборки пациентов, которую удаётся собрать, не достаточен для построения качественного классификатора. С другой стороны, способность препаратов ингибировать деление клеток тестируется на клеточных линиях, количество которых в десятки раз превосходит объем выборки пациентов, и для них также проводится количественный анализ генной экспрессии.

Данная работа посвящена разработке нового простого, интерпретируемого и быстрого метода, позволяющего использовать данные клеточных линий при прогнозировании результатов таргетной терапии. Для этого в работе применяются идеи монотонной классификации. Качество работы полученного классификатора было оценено на ретроспективных данных для 3 групп больных.

В главе 2 подробно описан алгоритм классификации, который мы предлагаем.

В главе 3 оценено качество работы алгоритма при прогнозировании результата лечения сорафенибом рака почки, лечения сорафенибом рака лёгкого и лечения паклитакселом рака груди. Для рака почки, где наш алго-

ритм показал наилучшие результаты и опередил методы из предыдущих работ, была проверена робастность классификации. Для рака груди наш классификатор не обеспечил хороший результат, поэтому для этой группы пациентов было решено параллельно с нашим алгоритмом опробовать другие методы машинного обучения.

2 Классификатор на основе множества детекторов

2.1 Описание данных

Классификатор для каждого из трёх рассматриваемых видов рака строился независимо. Для каждого вида рака имеется набор данных, где для каждого пациента предоставлен генный профиль и результат терапии (см. раздел "Использованные наборы данных"). Тех пациентов, которым терапия помогла, будем называть ответчиками, остальных – не ответчиками. По генному профилю каждого пациента при помощи алгоритма Oncofinder [6] были вычислены активации маркерных сигнальных путей (PAS) препарата (сорафениба для рака лёгкого и рака почки, паклитаксела для рака груди), которые и использовались как исходные признаки в представленном ниже алгоритме. Также использовался набор данных из 227 клеточных линий (см. раздел "Использованные наборы данных"). Для каждой клеточной линии, как и для пациентов, по генному профилю были посчитаны PAS маркерных сигнальных путей соответствующего препарата, таким образом, для каждого препарата клеточные линии, помощью которых оценивается его эффективность, и пациенты, которых лечили этим препаратом, описываются точками в одном и том же признаковом пространстве активаций маркерных сигнальных путей препарата. Тем не менее, для разных препаратов эти пространства различаются. Про каждую клеточную линию известно значение индекса IC50. Оно по определению равно концентрации лекарства, необходимой для снижения скорости деления клеток вдвое. Клеточные линии были упорядочены по возрастанию IC50 и разбиты квантилями 20%, 40%, 60% и 80% на 5 равных групп. Чтобы зашумленность измерений меньше влияла на результат исследования, вместо самого значения IC50 в работе использовался номер группы, к которой относится та или иная клеточная линия.

2.2 Описание детекторного классификатора

Определение 1 *Пусть f – некоторый признак (PAS некоторого сигнального пути). Пусть $p_{patients}(f)$ – вероятность того, что при случайном выборе 2-х пациентов из генерального распределения, один из которых ответчик, а второй – нет, окажется, что ответчик имеет большее значение f, чем не ответчик. В зависимости от того, $p_{patients}(f) > 0.5$ или*

$p_{patients}(f) < 0.5$, *будем называть признак* f *монотонно возрастающим на пациентах или монотонно убывающим на пациентах соответственно.*

Определение 2 *Пусть* $p_{cl}(f)$ *– вероятность того, что для двух случайно выбранных клеточных линий, относящихся к разным квантильным группам индекса IC50, клеточная линия с большим значениям индекса, также имеет большее значение признака* f*. Признаки* f*, для которых* $p_{cl}(f) > 0.5$ *назовём монотонно возрастающими на клеточных линиях, признаки, для которых* $p_{cl}(f) < 0.5$ *– монотонно убывающими на клеточных линиях.*

Предположение 1 *Монотонность пути не зависит от того, измерять её на пациентах, как это было сделано в определении 1, или на клеточных линиях, как это было сделано в определении 2.*

Предположение 1 основано на следующем рассуждении: если некоторый процесс (активность сигнального пути) позволяет лекарству лучше ингибировать деление клеточных линий, он, вероятно, позволит лекарству также лучше ингибировать деление раковых клеток. Это предположение позволяет применить элементы transfer learning, оценивая монотонность признаков по клеточным линиям и используя полученную информацию о монотонности признаков при классификации пациентов. Детекторный классификатор, описанный в данном разделе, основывается на предположении 1.

Далее предполагается наличие у нас в распоряжении некоторой меры монотонности μ. Интуитивно, $\mu(f, IC50)$ оценивает по выборке клеточных линий вероятность p из определения 2 для признака f, то есть чем выше $\mu(f, IC50)$, тем больше высокое значение f помогает препарату ингибировать деление клеток. На практике можно использовать самые разные меры монотонности μ. Будем считать, что $\mu(f, IC50) > 0$ для возрастающих по клеточным линиям признаков и $\mu(f, IC50) < 0$ для убывающих по клеточным линиям признаков. Выбор μ описан в разделе "Выбор меры монотонности".

Основной идеей нашего алгоритма является построение множества детекторов ответчиков и детекторов не ответчиков, на основании которых и делается итоговая классификация. Эти детекторы строятся по четырем правилам:

1. Для каждого **ответчика** r из обучающей выборки пациентов и **возрастающего** на клеточных линиях признака f строится детектор **ответчиков**, который срабатывает на пациентах p таких, что $f(p) > f(r)$.

2. Для каждого **ответчика** r из обучающей выборки пациентов и **убывающего** на клеточных линиях признака f строится детектор **ответчиков**, который срабатывает на пациентах p таких, что $f(p) < f(r)$.

3. Для каждого **не ответчика** n из обучающей выборки пациентов и **возрастающего** на клеточных линиях признака f строится детектор **не ответчиков**, который срабатывает на пациентах p таких, что $f(p) < f(n)$.

4. Для каждого **не ответчика** n из обучающей выборки пациентов и **убывающего** на клеточных линиях признака f строится детектор **не ответчиков**, который срабатывает на пациентах p таких, что $f(p) > f(n)$.

Алгоритм классификации на основе детекторов, которые мы предлагаем, имеет следующий вид:

1. Определить возрастающие признаки inc_i и убывающие признаки dec_i по клеточным линиям. Монотонность здесь определяется знаком $\mu(f, IC50)$.

2. Найти главные компоненты inc_i^{pc} $(i = 1, 2, ..., |\mathrm{inc}|)$ проекции выборки клеточных линий на $\mathrm{span}\,(\mathrm{inc})$ и главные компоненты dec_i^{pc} $(i = 1, 2, ..., |\mathrm{dec}|)$ проекции выборки клеточных линий на $\mathrm{span}\,(\mathrm{dec})$. Здесь $\mathrm{span}(X)$ обозначает линейную оболочку множества X в пространстве признаков. Количество главных компонент равно количеству возрастающих и убывающих признаков соответственно, то есть понижения размерности здесь не происходит.

3. Для каждого из признаков $f \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}$ вычислить $m(f) = \mu(f, IC50)$.

4. По обучающей выборке пациентов построить все возможные детекторы ответчиков и не ответчиков на признаках $f \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}$ по четырем описанным выше правилам. Здесь, как и раньше, монотонность определяется в зависимости от значения $\mathrm{sign}(m(f))$

5. Вычислить веса признаков

$$w(f) = \frac{|m(f)|}{\sum_{f' \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}} |m(f')|} \qquad \forall f \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}$$

   Веса детекторов вычисляются по весам признаков следующим образом: для детектора d вес $w(d) = w(f_d)/n$, где f_d – признак, по которому был построен детектор d, n – количество пациентов пациентов того класса, который детектирует d, в обучающей выборке.

6. Для каждого пациента p из тестовой выборки следует определить сработавшие детекторы. r-значение вычисляется как суммарный вес сработавших детекторов ответчиков, n-значение – суммарный вес сработавших детекторов не ответчиков. Вероятность принадлежности p классу ответчиков оценивается как $\frac{r}{r+n}$. В зависимости от того, превышает ли эта вероятность порог τ, происходит классификация.

Использование метода главных компонент [7] в пункте 2 алгоритма и построение детекторов на inc^{pc}, dec^{pc}, а не на изначальных признаках inc, dec связано с тем, что наш алгоритм чувствителен к выбору базиса. Метод главных компонент показался нам наиболее естественным способом выбора базиса, и детекторы, построенные на главных компонентах, в совокупности дали более высокое качество классификации, чем детекторы, построенные

непосредственно на активациях сигнальных путей. Возможно, это отчасти связано с тем, что главные компоненты, в отличие от изначального набора признаков, представляют собой ортонормированный базис, то есть их ковариационная матрица тождественна. Искать главные компоненты непременно надо отдельно в подпространстве возрастающих признаков и отдельно – в подпространстве убывающих признаков, т.к. иначе построенные компоненты будут содержать в себе случайную смесь из возрастающих и убывающих признаков, и строить на них пороговые детекторы будет бессмысленно.

В разделе "От отбора признаков к весам детекторов"дана мотивация для использования весов детекторов в пункте 5 алгоритма.

Можно отметить ещё одну черту детекторного классификатора: каждый из его детекторов использует только один признак. Таким образом, многомерность данных не используется при построении каждого детектора в отдельности, но используется для увеличения размера ансамбля детекторов. Тут можно увидеть сходство нашего метода с наивным байесовским классификатором, где многомерная задача вычисления правдоподобия выборки сводится к набору одномерных задач. Однако, в отличие от наивного байесовского классификатора, детекторный классификатор опирается на монотонность признаков, за счёт чего удаётся перейти от оценок плотностей одномерных распределений к оценкам функций одномерных распределений, что, фактически, и происходит при усреднении всех детекторов одного класса, построенных по одной и той же компоненте. То, что каждый детектор использует только один признак, также положительно сказывается на интерпретируемости классификации.

2.3 Выбор меры монотонности

В этом разделе предложены некоторые меры монотонности $\mu : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, которые можно использовать для определения монотонности признаков и весов детекторов. Неформально, все предложенные меры μ по наборам чисел X, Y оценивают, насколько часто $x_i > x_j \Rightarrow y_i > y_j$. Если это следствие справедливо для всех i, j, назовём пару (X, Y) монотонной относительно μ.

1. Рассмотрим перестановку π такую, что $(\pi(X), Y)$ – монотонная пара. Пусть t – количество инверсий в перестановке π, $|X| = |Y| = n$. Тогда доля инверсий в перестановке π равна $\frac{2t}{n(n-1)}$. Мера монотонности $\mu_1 = 1 - \frac{4t}{n(n-1)}$.

2. Коэффициент корреляции Пирсона ρ. Здесь X, Y рассматриваются как дискретные случайные величины. Мера монотонности $\mu_2 = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}}{\sqrt{\mathbb{D}(X)\mathbb{D}(Y)}}$.

3. Коэффициент монотонности ρm [2]. Здесь X, Y также рассматриваются как дискретные случайные величины. Пусть F_X – функция распределения случайной величины X, F_X^{-1} – обратная ей функция, U – равномерно распределённая на $[0, 1]$ случайная величина, $X^* = F_X^{-1}(U)$,

$X' = F_X^{-1}(1-U)$.

$$\mu_3(X,Y) = \rho m(X,Y) = \begin{cases} \frac{\mathrm{Cov}(X,Y)}{\mathrm{Cov}(X^*,Y^*)} & \mathrm{Cov}(X,Y) > 0 \\ 0 & \mathrm{Cov}(X,Y) = 0 \\ -\frac{\mathrm{Cov}(X,Y)}{\mathrm{Cov}(X^*,Y')} & \mathrm{Cov}(X,Y) < 0 \end{cases}$$

Хотя интуитивно μ_1 лучше всего отражает идею монотонности, более качественные модели удалось построить с μ_2 и μ_3. Результаты для этих двух мер получились практически идентичными, и сравнивать их по какой-либо метрике оказалось бессмысленно. При этом μ_2 проще и популярней как мера монотонности, в то время как μ_3 всегда совпадает с μ_2 по знаку, а также обладает рядом свойств, которые естественно ожидать от меры монотонности. Например, если X, Y – выборка из пары комонотонных случайных величин, $\rho m(X,Y) = 1$, а если X, Y – выборка из пары контрмонотонных случайных величин, $\rho m(X,Y) = -1$. В этом случае построенные по такому признаку детекторы никогда не будут ошибаться, хотя по значению обычного коэффициента корреляции Пирсона понять это не всегда возможно: он может не равняться ± 1.

В следующей главе все результаты приведены для меры μ_2.

3 Тестирование детекторного классификатора

3.1 Выбор способа оценки качества классификации

Как часто врачи ошибаются при назначении терапии, неизвестно, поэтому точный критерий применимости классификатора для выбора курса лечения не выработан. Тем не менее, можно сформулировать некоторые необходимые свойства, которыми должен обладать такой классификатор. Они позволяют на интуитивном уровне оценить применимость метода. От классификатора, который мы строим, требуется, в первую очередь, верно определять больных, которым данная терапия не поможет (не ответчиков), потому что ошибка классификатора на не ответчике может немедленно привести к неверному выбору курса лечения. В частности, хороший классификатор должен иметь $FPR = \frac{FP}{N} < 0.1$, где FP – количество не ответчиков, классифицированных как ответчики, а N – количество всех не ответчиков в наборе данных. Аналогично определяется FN – количество ответчиков, классифицированных как не ответчики, P – количество всех ответчиков в наборе данных, $FNR = \frac{FN}{P}$. Ограничение на FNR для классификации в данном случае слабое. Приблизительное требование, которое здесь предъявляется к классификатору, – $FNR < 0.5$. Действительно, даже если классификатор забракует некоторое лекарство для потенциального ответчика, это не помешает найти подходящее среди десятков других лекарств.

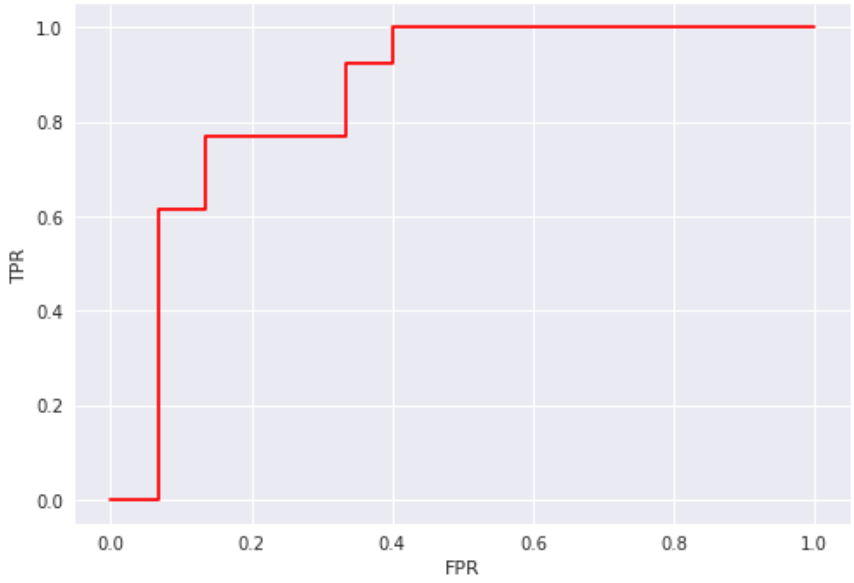
Если в наборе данных R ответчиков и N не ответчиков, то FPR и FNR могут принимать только значиния вида $\frac{x}{N}$ и $\frac{y}{R}$ сответственно, где x, y – целые неотрицательные числа. При малых значениях R, N, сравнение клас-

сификаторов по этим метрикам часто будет бессмысленным. В течение последних десятилетий в медицинской диагностике активно использовалась другая метрика – ROC AUC [3], [4]. Значение ROC AUC равно вероятности того, что при случайном выборе одного ответика и одного не ответчика из выборки классификатор присвоит ответчику большую вероятность успеха терапии, чем не ответчику. Эта метрика одна из наиболее часто используемых в машинном обучении, см., например, классическую работу [5]. В работе [1], где решается та же задача, что и в настоящей работе, классификаторы также сравнивались по метрике ROC AUC. По этим причинам в качестве метрики качества модели было решено использовать именно ROC AUC. Такой подход подразумевает, что сначала строится регрессор $f : X \to \mathbb{R}$, итоговый классификатор имеет вид $\sigma_\tau(f(x))$, где $\sigma_\tau(z) = 1$, если $z > \tau$ и 0 иначе. Значение τ подбирается отдельно.

Основным способом оценки качества классификации была выбрана Leave-One-Out кросс-валидация. С её помощью для каждого из пациентов была оценена вероятность принадлежности классу ответчиков. При этом детекторы каждый раз строились по всем пациентам кроме того, для которого делался прогноз. Поскольку истинные классы пациентов известны, по всем полученным прогнозам можно построить ROC кривую и найти площадь под ней – это и будет искомое значение метрики ROC AUC.

3.2 Лечение рака почки сорафенибом

Способность детекторного классификатора прогнозировать результат терапии рака почки сорафенибом была проверена на наборе данных, предоставленном в работе [1] (см. раздел "Использованные наборы данных"). Этот набор состоит из геных профилей 28 больных раком почки, которых лечили сорафенибом. 13 больных ответили на терапию, 15 – нет. Значение ROC AUC оценивалось с помощью процедуры Leave-One-Out, в соответствии с описанием из предыдущего раздела. Полученная ROC кривая представлена ниже, площадь под ней равна 0.86:

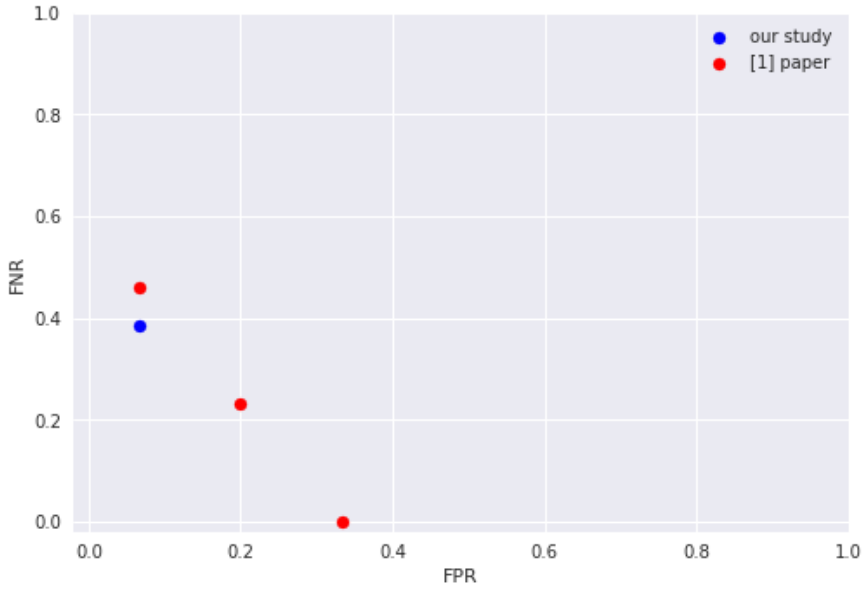


При вычислении активаций сигнальных путей необходимо выбрать нормировку. В данной работе использовалась нормировка на глиальную ткань. В статье [1] на этих же массивах больных и клеточных линий с нормировкой на глиальную ткань было получено значение ROC AUC = 0.77. Также в статье [1] использовались и другие варианты нормировки, которые давали на данных выборках ROC AUC = 0.81 и 0.82 соответственно. В любом случае, ROC AUC = 0.86 – улучшение качества классификации по сравнению с [1]. То, что удалось улучшить результат работы [1], неожиданно, т.к. в этой работе использовался существенно более сложный классификатор. В работе [1] представленные значения достигались при подобранных по сетке гиперпараметрах, причем выбор значений гиперпараметров производился по всей выборке. Таким образом, оценки качества получались завышенными. Некоторая робастность относительно выбора гиперпараметров в работе [1] доказана, однако смещенности оценки ROC AUC это не отменяет.

Также можно отметить, что детекторный классификатор работает очень быстро при объёме выборок в десятки пациентов и сотни клеточных линий: Leave-One-Out предсказание для всех пациентов занимает 1 секунду. Для сравнения, модель, предложенная в работе [1], делает такое предсказание примерно за полчаса. Время работы алгоритма – не самый важный фактор в нашей задаче, однако, если использовать детекторные классификаторы как часть более сложной модели, высокая скорость работы может стать важным фактором.

Модель, обладающая ROC кривой, представленной выше, способна верно определять 62% (или 8 из 13) ответчиков и 93% (14 из 15) не ответчиков, т.е. $FPR = \frac{1}{15}$, $FNR = \frac{5}{13}$. Такие значения достигаются при пороге классификации $\tau = 0.53$. Сравнить эти значения с результатами работы [1] на

трёх различных нормировках можно в следующем графике.



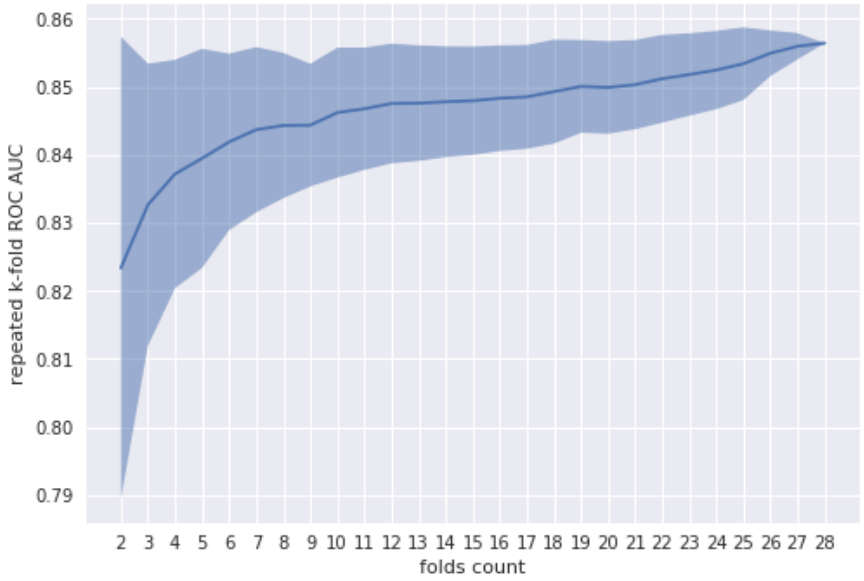
Поскольку набор данных пациентов использовался небольшой, для того, чтобы убедиться в эффективности прогноза для рака почки, необходимо проверить робастности модели.

3.3 Проверка робастности прогноза для рака почки

Детекторный классификатор обучается на двух выборках: выборке клеточных линий и обучающей выборке пациентов. В данном разделе представлены результаты тестирования робастности классификации относительно изменения обеих выборок.

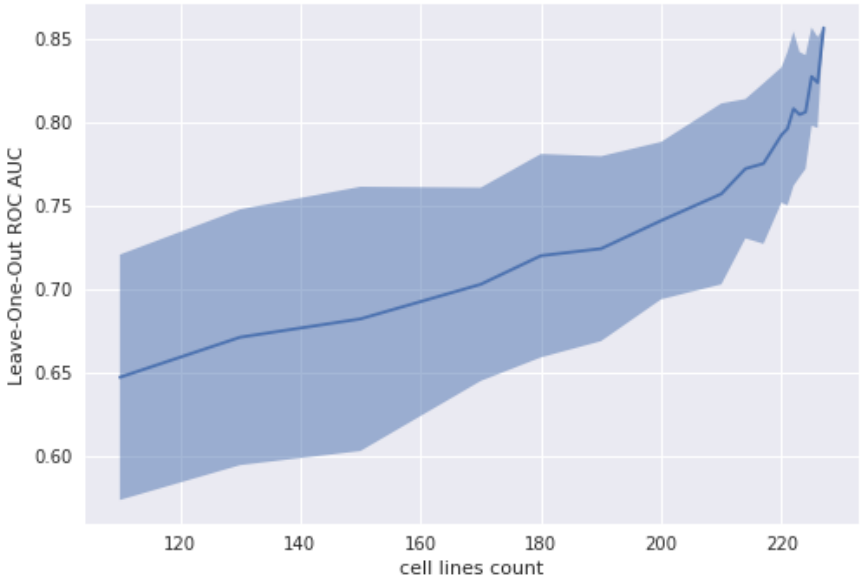
Чтобы проверить робастность относительно обучающей выборки пациентов, было использовано многократное повторение процедуры кросс-валидации k-fold для разного количества фолдов. Для каждого фиксированного количества фолдов k выборка пациентов 100 раз разбивалась на k групп, отличающихся по мощности не более, чем на 1. Далее, по очереди каждая из групп была использована как тестовая выборка, а объединение остальных $k-1$ групп – как обучающая выборка пациентов. В итоге получилось 100 наборов оценок вероятности принадлежности классу ответчиков для всех пациентов. Для каждого такого набора было вычислено значение ROC AUC. При $k = 28$ все 100 полученных значений соответствуют значению ROC AUC, которое получается при кросс-валидации Leave-One-Out. При $k = 14$, в отличие от Leave-One-Out, каждый пациент классифицируется на основе детекторов построенных не по всем 27 остальным пациентам, а только по 26 пациентам, то есть объем каждой обучающей выборки уменьшился

на 1. $k = 2$ – предельный случай, когда мощность обучающей выборки снижается вдвое по сравнению с Leave-One-Out. Естественно, снижение объема обучающей выборки влечет за собой менее точный прогноз и качество классификации падает. На графике ниже для каждого k от 2 до 28 показано среднее значение ROC AUC по 100 разбиениям для данного k. Также вокруг среднего значения изображен интервал $\pm$ стандартное отклонение на тех же 100 разбиениях.



В целом можно сказать, что классификация стабильна относительно выборки пациентов: даже в случае 2-fold, когда обучающая выборка пациентов уменьшилась вдвое, среднее значение ROC AUC = 0.82, то есть качество классификации остаётся хорошим.

Посмотрим теперь, что происходит при изменении выборки клеточных линий. В детекторном классификации клеточные линии используются для определения монотонности признаков и подсчёта весов детекторов. Попробуем делать это не по всем 227 клеточным линиям, а только по части из них. Будем выбирать случайное подмножество определенной мощности из выборки клеточных линий и использовать для обучения только их, причем будем повторять каждый такой эксперимент 50 раз, используя разные подмножества клеточных линий определённой мощности. Зависимость ROC AUC от количества используемых линий изображена на графике ниже. Так же, как и на предыдущем графике, изображено стандартное отклонение в виде доверительного интервала.



Как и ожидалось, если использовать не все клеточные линии, оценки монотонности признаков становятся менее точными, а веса детекторов – более зашумленными, поэтому качество модели падает: например, если использовать только 110 клеточных линий, среднее значение ROC AUC получается равным 0.65, а если использовать 220 клеточных линий, то среднее значение ROC AUC уже 0.79. Неожиданно большой оказалась разница между исползованием 220 и 227 клеточных линий. Существенный перепад можно объяснить тем, что данная выборка из 227 клеточных линий оказалась удачной для своего объёма. При построении графика производилось усреднение по 50 различным выборкам из 220 клеточных линий, и качество эти выборки давали разное, в том числе иногда получались результаты лучше, чем на всех 227 пациентах (ROC AUC до 0.91). Так что значение ROC AUC = 0.86 на 227 клеточных линиях не является выбросом.

Из этого эксперимента можно сделать вывод, что массив клеточных линий и подобранные по нему веса детекторов играют ключевую роль в представленном классификаторе. В целом, увеличение массива клеточных линий существенно увеличивает качество классификации. Поэтому один из способов улучшить классификатор – увеличить массив клеточных линий или в рамках данного объема выборки постараться отобрать лучшие клеточные линии для оценки по ним весов (например, удаляя выбросы при оценке монотонности каждого признака).

3.4 Лечение рака лёгкого сорафенибом

Монотонность влияния активаций сигнальных путей на вероятность успеха терапии согласно нашему алгоритму определяется по выборке клеточных

линий, без использования самих пациентов. В данном разделе рассматривается возможность применения классификатора на основе детекторов для случая лечения сорафенибом рака лёгкого. Поскольку в применении к раку почки также изучался сорафениб, сейчас клеточные линии можно использовать те же, что и для рака почки, и в таком случае монотонности компонент и веса детекторов не изменятся, отличаться будут только сами детекторы и классифицируемые пациенты.

Мы использовали набор данных GSE31428 из репозитория GEO (см. раздел "Использованные наборы данных"). В нём представлена информация (генные профили и результат терапии) о 37 пациентах с раком лёгкого, которых лечили сорафенибом. К этим больным был применён детекторный классификатор с помощью процедуры кросс-валидации Leave-One-Out, как это делалось раньше. Был выбран порог классификации $\tau = 0.5$. В итоге метрики качества классификации получились такими: ROC AUC = 0.63, FPR=0.36, FNR = 0.35. Таким образом, для рака лёгкого результаты получились хуже, чем в статье [1], где с разными нормировками были получены значения ROC AUC = 0.72, 0.77, 0.78.

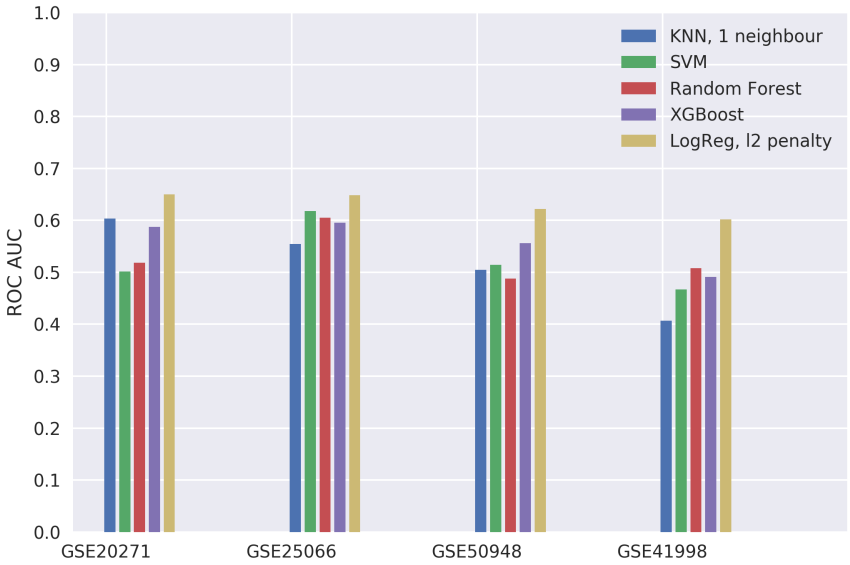
На этом же наборе данных были протестированы самые распространённые алгоритмы машинного обучения, такие как метод ближайших соседей, Random Forest, XGBoost, логистическая регрессия, SVM. Качество оценивалось, как и раньше, с помощью процедуры Leave-One-Out. Оказалось, что ROC AUC для этих моделей не превосходит 0.5, т.е. эти методы вообще не работают на нашем наборе данных пациентов. Таким образом, в сравнении с ними наш метод переноса информации с кеточных линий на пациентов дал существенное улучшение.

3.5 Лечение рака груди паклитакселом

Данный раздел посвящён описанию тестирования нашего классификатора в невыгодных для него условиях. На этот раз эксперимент проводился для больных раком груди, которых лечили паклитакселом. При подсчёте активаций сигнальных путей производилась нормировка экспрессий на среднее геометрическое по выборке.

Эксперимент проводился на наборах данных GSE41998, GSE25066, GSE20271 и GSE50948 из репозитория GEO. Эти наборы содержат от 84 до 508 пациентов. ROC AUC детектроного классификатора на этих наборах данных принимает значения от 0.44 до 0.55. Таким образом, наш классификатор на них не работает. Как было показано в разделе 3.3, наш классификатор не очень чувствителен к изменениям выборки пациентов, поэтому он не смог в полной мере воспользоваться приемуществом большой выборки пациентов, в то время как изменение признакового пространства оказалось для него очень важным: в пространстве маркерных сигнальных путей сорафениба классификатор как для рака почки, так и для рака лёгкого работал существенно лучше, чем в пространстве маркерных сигнальных путей паклитаксела.

Для таких больших наборов данных, похоже, перенос данных с клеточных линий не нужен. Ниже приведена диаграмма, на которой сравнены несколько методов машинного обучения (с подобранными оптимальными значениями гиперпараметров) на этих данных. Все классификаторы обучались только по массиву пациентов, клеточные линии не использовались. Лучше всего показала себя логистическая регрессия с регуляризацией l_2. Эта модель на всех 4-х наборах данных по раку груди имела ROC AUC от 0.6 до 0.65.



4 От отбора признаков к весам детекторов

В этой главе предложено семейство классификаторов, совсем немного отличающихся от алгоритма классификации на основе детекторов, описанного в разделе 2.2. Сами по себе они классифицируют представленные наборы данных хуже, чем описанный ранее классификатор, зато открывают возможности для дальнейших исследований и могут служить демонстрацией, объясняющей происхождение весов детекторов в алгоритме из раздела 2.2. Идея заключается в том, что зависимость вероятности ответа на терапию от некоторых признаков из нашего пространства может быть существенно не монотонной. Такие признаки могут негативно влиять на качество работы классификатора, поэтому их, возможно, стоит отбрасывать. Обзаведёмся

пороговым значением θ и будем считать, что если $\mu_2(f, IC50) > \theta$, то признак возрастающий, если $\mu_2(f, IC50) < -\theta$, то признак убивающий, а если $-\theta < \mu_2(f, IC50) < \theta$, то признак не монотонный и при построении пороговых детекторов его не стоит учитывать.

Классификатор c_{θ_1,θ_2}, реализующий эту идею, выглядит следующим образом:

1. Определить возрастающие признаки inc_i и убывающие признаки dec_i по клеточным линиям. Монотонность здесь определяется сравнением $\mu_2(f, IC50)$ с $\pm\theta_1$, как описано выше. Из-за наличия порога $\theta_1 > 0$, на этом шаге мы отбросим те признаки, которые с точки зрения меры μ не монотонные.

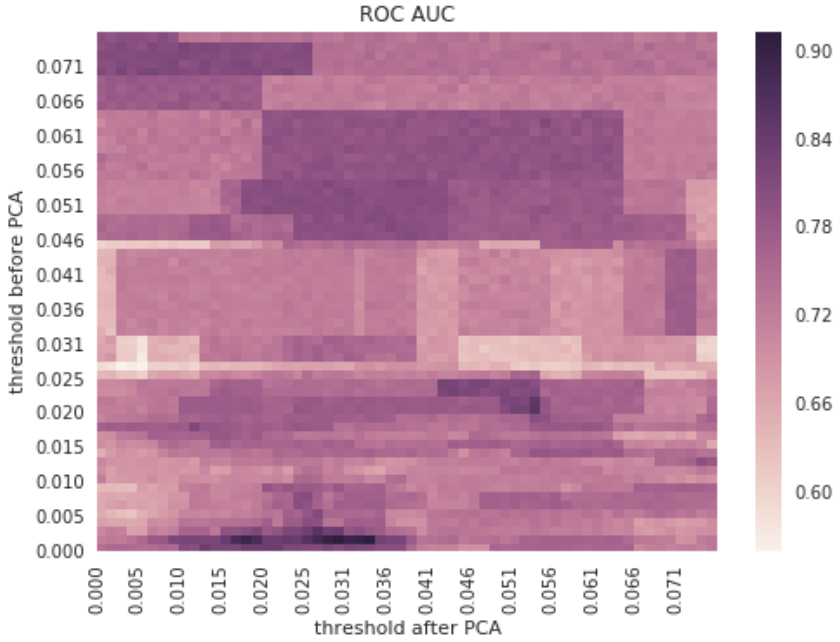
2. Найти главные компоненты inc_i^{pc} и dec_i^{pc} проекции выборки клеточных линий на span (inc) и span (dec) соответственно. Количество главных компонент здесь по-прежнему равно размерности пространства, то есть понижения размерности не происходит.

3. Из главных компонент подпространства возрастающих признаков inc^{pc} и главных компонент подпространства убывающих признаков dec^{pc} отбросим не монотонные компоненты f, для которых $|\mu_2(f, IC50)| < \theta_2$.

4. По обучающей выборке пациентов построить все возможные детекторы ответчиков и не ответчиков по прведённым в разделе 2.2 правилам 1-4 для признаков $f \in \text{inc}^{pc} \cup \text{dec}^{pc}$.

5. Для каждого пациента p из тестовой выборки следует определить сработавшие детекторы. r-значение вычисляется как доля сработавших детекторов среди всех детекторов ответчиков, n-значение – как доля сработавших среди детекторов не ответчиков. Вероятность принадлежности p классу ответчиков оценивается как $c_{\theta_1,\theta_2}(p) = \frac{r}{r+n}$. В зависимости от того, превышает ли эта вероятность порог τ, происходит классификация.

Как и раньше, оценим ROC AUC с помощью кросс-валидации Leave-One-Out по выборке больных. На тепловой карте ниже представлено значение ROC AUC классификатора c_{θ_1,θ_2} на наборе данных по раку почки в зависимости от гиперпараметров θ_1 (по вертикали) и θ_2 (по горизонтали).



При абсолютно любых значениях гиперпараметров θ_1, θ_2 имеем ROC AUC > 0.5, среднее значение ROC AUC по всем точкам на представленной тепловой карте равно 0.74, максимальное значение равно 0.91. Высокое среднее значение на тепловой карте означает, что, несмотря на то, что классификаторы c_{θ_1,θ_2} с разными значениями гиперпараметров θ_1, θ_2 сильно скоррелиррованы, ансамбль из этих классификаторов будет иметь высокое качество. Также видно, что наилучшие значения достигаются при $\theta_1 \approx 0$, поэтому в ансамбль будем включать только классификаторы вида c_{0,θ_2}. Чем больше детекторов усредняет классификатор c_{0,θ_2}, тем он робастней, поэтому вес классификатора c_{0,θ_2} в ансамбле положим равным количеству детекторов, входящих в него.

Пусть $d(f, p)$ – детектор класса пациента p, построенный по признаку f и пациенту p, $c(f)$ – классификатор, построенный по всем детекторам вида $d(f, p_i)$ для данного признака f и всех больных p_i из выборки. Положим также

$$S_{\theta_2} = \{f \in \text{inc}^{pc} \cup \text{dec}^{pc} : |\mu_2(f, IC50)| > \theta_2\}$$

Тогда

$$c_{0,\theta_2} = \frac{1}{|S_{\theta_2}|} \sum_{f \in S_{\theta_2}} c(f)$$

И ансамбль имеет вид

$$\int_{\theta_2=0}^{\max_f |\mu_2(f,IC50)|} |S_{\theta_2}| c_{0,\theta_2} = \int_{\theta_2=0}^{\max_f |\mu_2(f,IC50)|} \sum_{f \in S_{\theta_2}} c(f) =$$

$$\sum_{f \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}} \int_0^{|\mu_2(f, IC50)|} f(c) = \sum_{f \in \mathrm{inc}^{pc} \cup \mathrm{dec}^{pc}} |\mu_2(f, IC50)| f(c)$$

В итоге ансамбль из классификаторов, лежащих в нижней строке представленной выше тепловой карты, оказался эквивалентен детекторному классификатору из раздела 2.2. Такой переход от двухпараметрического семейства классификаторов к их ансамблю не только повышает ROC AUC, но и избавляет алгоритм от гиперпараметров, подбор которых по выборке из 28 пациентов был бы затруднителен. Для дальнейшего улучшения классификации можно ограничить θ_2 при составлении ансамбля, однако это потребует введение дополнительных гиперпараметров, что повлияет на обобщающую способность модели.

5 Заключение

Представленный в работе детекторный классификатор на основе детекторов неожиданно хорошо показал себя для прогнозирования терапии рака почки сорафенибом: качество классификации превосходит даже результат более сложной модели из предыдущей работы [1], причём классификация устойчивая к изменению выборки пациентов. Этот же классификатор можно использовать при прогнозировании результата терапии других онкологических заболеваний сорафенибом, но качество классификации, похоже, получается ниже, чем для рака почки.

Для рака груди детекторный классификатор не сработал. Это, вероятно, связано с тем, что данные по раку груди лежали в совершенно ином пространстве, чем данные по раку почки и лёгкого. Помимо того, что препараты и, следовательно, набор маркерных сигнальных путей были разными, для рака груди и паклитаксела мы не располагали набором данных для нормировки в алгоритме Oncofinder [6] на нейроглиальные клетки, как мы это делали для сорафениба. Поэтому нормировку в Oncofinder для рака груди пришлось делать на среднее геометрическое по выборке пациентов. Отметим, что при нормировке на среднее геометрическое вместо нормировки на нейроглиальные клетки показатели для сорафениба также падали до ROC AUC = 0.68 вместо 0.86 для рака почки и ROC AUC = 0.60 вместо 0.63 для рака легкого. Таким образом, естественный способ построить работающий детекторный классификатор для паклитаксела – найти подходящие данные для нормировки на глиальную ткань.

Ещё один способ улучшить качество классификации – подобрать подходящий массив клеточных линий. Как было показано в разделе 3.3, детекторный классификатор довольно чувствителен к этому выбору. Возможным представляется два способа улучшения массива клеточных линий. Первый способ – увеличение объема выборки клеточных линий для более точной оценки монотонности признаков и весов детекторов. Второй способ – отбор наиболее релевантных клеточных линий, поскольку, как было показано

в разделе 3.3, набор клеточных линий для рака почки оказался довольно удачным, но даже для него удаление лишних клеточных линий может увеличить качество классификации. Для паклитаксела, похоже, набор клеточных линий оказался неудачным, но, отбросив часть из них, скорее всего, можно получить хороший классификатор. Поэтому встаёт вопрос о методе отбора подходящих клеточных линий. Разработка такого метода представляется ключевым шагом дальнейшего исследования.

Использованные наборы данных

1. Генные профили и результаты терапии для больных раком почки опубликованы в статье [1]
2. Нормировочный массив для больных раком почки
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49972
3. Генные профили и результаты терапии для больных раком легкого
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31428
4. Нормировочный массив для больных раком легкого
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43458
5. генные профили больных раком груди:
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41998,
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25066,
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20271
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50948
6. Генные профили клеточных линий
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68950
7. Нормировочные масссивы для клеточных линий:
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14805,
   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21212
8. Индекс IC50 для клеточных линий
   https://www.cancerrxgene.org/translation/Drug/30

Реализация метода

Реализация описанного в данной работе метода на языке Python 3:
https://github.com/Mixanik-43/Detector-Based-Classifier

Список использованной литературы

1. N. Borisov et al. — *A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency.* 2018. Cell Cycle, 17(4):486-491, DOI: 10.1080/15384101.2017.1417706

2. Farida Kachapova et al. — *A Measure of Monotonicity of Two Random Variables.* 2012. Journal of Mathematics and Statistics 8 (2): 221-228

3. Mark H. Zweig et al. — *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.* 1993. Clinical Chemistry 39 (8): 561–577.

4. Pepe Margaret S. — *The statistical evaluation of medical tests for classification and prediction.* 2003. New York, NY: Oxford

5. Spackman et al. — *Signal detection theory: Valuable tools for evaluating inductive learning.* 1989. Proceedings of the Sixth International Workshop on Machine Learning: 160–163, San Mateo, CA: Morgan Kaufmann.

6. A. Buzdin et al. — *Oncofinder, a new method for the analysis of intracellular signaling pathway activation using transcriptomic data.* 2014. Front Genet. DOI:10.3389/fgene.2014.00055. PMID:24723936

7. Pearson K. — *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, (1901) 2, 559—572

8. Ivan V. Ozerov et al. — *In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development.* 2016. Nature communications, DOI: 10.1038/ncomms13427, PMID: 27848968