Использование метода понижения размерности в медицинских исследованиях.

2 июня 2017 г.

1 Биологическая постановка задачи

Одной из наиболее важных задач, возникающих при лечении онкозаболевания, является предсказание успешности терапии, предлагаемой пациенту. Важность данной задачи сложно переоценить, потому как, зачастую, подходы, используемые для борьбы с онкологическими заболеваниями, являются не только крайне дорогосоящими, но и не безвредными для организма человека. Благодаря новым технологиям, сегодня достаточно даже небольшого кусочка ткани, чтобы получить полную картину о геноме пациента. Известно, что активность генома меняется в зависимости от типа ткани и что, в частности, она специфична для ткани, соответствующей опухоли. Таким образом, подобные данные не только несут в себе крайне важную информацию о состоянии здоровья пациента, но и могут быть использованы для решения задачи о предсказании эффективности лечения.

Для решения задач такого типа подразумевается наличие испытуемой группы пациентов, для каждого из которых составлен генетический профайл активности генома в опухоли (иными словами, измерены его *генные экспрессии*). Кроме того, необходимо иметь информацию об успешности лечения данного типа рака данным типом терапии. В настоящей работе исследовалась успешность *таргетной терапии*, как подхода к лечению *миеломы*. В качестве испытуемой группы выступили пациенты из международной базы данных.

2 Описание данных

Данные, участвующие в исследовании, содержат информацию о 250 пациентах, из которых 11 человек не болели миеломой. Среди оставшихся 239 пациентов, больных миеломой, 113 человек откликнулись на таргетную терапию, а 126 - нет. О каждом из пациентов известны значения 13832 экспрессий. Так как 11 здоровых

пациентов в базе данных не получали никакого лечения, то они не могут быть использованы для предсказания успешности терапии, отвечающего успешности таргетной терапии, поэтому исключим их из рассмотрения.

Данная биологическая задача может быть переформулирована в терминах машинного обучения следующим образом. В качестве *факторов* будем рассматривать величины генных экспессий, а *отклику* будет соответствовать факт успешности/не успешности терапии. Таким образом, имеется матрица X размерности 239×13832, в i-той строке которой записаны значения генных экспрессий i-того пациента. Вектор Y составляется из булевых значений, где 1 на i-том месте вектора соответствует тому, что i-тый пациент откликнулся на таргетную терапию, а 0 кодирует неуспешность лечения. Задача заключается в построении классификатора, который по данной матрице X совершает наиболее точное предсказание вектора откликов Y.

3 Стратегия исследования, описание экспериментов

Основной проблемой построения классификатора на полученных данных является размерность полученных данных. С одной стороны, так как количество факторов в почти 60 раз превосходит объем выборки, велик факт возникновения переобучения. С другой стороны, большое количество факторов приводит к чрезмерным вычислительным затратам при построении классификатора. Помимо проблем, возникающих в машинном обучении, понижение размерности данных может быть полезно и с точки зрения биологических приложений. Уменьшение размерности данных позволяет редуцировать количество информации, необходимой для назначения лечения пациенту. Так как на составление генетического профайла активности генома уходит некоторое существенное количество времени и ресурсов, решение данной задачи позволило бы существенно сократить расходы медицинского центра и ускорить процесс назначения лечения.

Таким образом, одним из этапов исследования, предшествующих непосредственному построению классификатора, является сокращение размерности данных. В данном исследовании рассматривалось два различных метода сокращения размерности. Первый применяемый метод — метод главных компонент, который заключается в поиске подпространства меньшей размерности такого, что разброс ортогональных проекций данных на это пространство максимален.

В основе второго подхода лежит метод экстремальной группировки. Суть данного метода состоит в разбиении факторов на группы такие, что факторы из одной группы коррелируют между собой, а факторы из разных групп имеют слабую корреляцию. Параметром данного метода является количество групп, на которые разбиваются факторы. Помимо разбиения на группы алгоритм метода экстремаль-

ной группировки выдает набор *центроидов* для каждой из полученных групп. Центроидом группы называется случайная величина, которая наиболее сильно коррелирована с факторами, попавшими в данную группу. Второй метод понижения размерностей использует центроиды каждой из групп, построенных методом экстремальной группировки, в качестве признаков для построения классификатора. Кроме того, будет рассмотрено несколько подходов к выбору среди всех центроидов некоторого поднабора "оптимальных".

Второй частью исследования является построение классификатора. В данной работе рассматривалось два основных метода классификации: метод ближайших соседей и метод опорных векторов. В качестве параметра первого метода выступает количество соседей(от 3 до 15), а второго - ядро (линейное, полиномиальное, радиальное и сигмоидальное) и параметры ядра.

Классификация строится двумя способами. Первый - непосредственное построение классификатора на основе матрицы признаков (до или после понижения размерности). Вторая классификация опирается на группы, полученные в методе экстремальной группировки, и метод голосования. Более подробно, предсказание отклика строится следующим образом: каждая группа строит предсказание отклика на основе тех факторов, которые попали в данную группу. Если более половины групп предсказали положительный отклик (соответствующий единице), то результатом классификации будем считать 1. В противном случае будем считать, что предсказываемый отклик - отрицательный (соответствующий 0).

Качество предсказания будет оцениваться несколькими способами. Наиболее важной характеристикой с точки зрения биологической постановки задачи является процент правильно предсказанных отрицательных откликов (specifisity). Неправильное предсказание этого класса означает, что то (иногда не безвредное) лечение, которое получит пациент, окажется неэффективным и не даст никаких результатов. Вторая важная характеристика - число правильно предсказанных положительных классов. В качестве третьей метрики мы будем рассматривать менее интуитивную величину aucroc, равную площади под roc-кривой, которая отражает зависимость между долей верных классификаций и долей ложных классификаций при изменении порога решающего правила.

Основной целью исследования является сравнение вышеизложенных методов с точки зрения одной из характеристик качества предсказания. Кроме того, интересным вопросом является зависимость методов, основанных на методе экстремальной группировки от количества групп.

4 Описание результатов

4.1 Построение классификатора на основе всех экспрессий

В результате предварительной обработки данных было выяснено, что в матрице X есть повторяющиеся столбцы. Так как эти столбцы не несут никакой дополнительной информации с точки зрения построения классификатора, удалим эти 626 столбцов из матрицы X.

Перед тем, как применять различные методы понижения размерности, построим классификатор на основе всех 13206 признаков-экспрессий. В качестве первого метода классификации используем метод ближайших соседей KNN (далее N будет обозначать количесво соседей). Так как параметр N может сильно влиять на качество классификатора, рассмотрим различные значения этого параметра. А именно, пусть количество соседей меняется в пределах от 3 до 15. Для каждого фиксированного N вычислим следующие характеристики качества классификации:

- *accuracy score* — количество правильно предсказанных откликов;
- *0-class score* — количество правильно предсказанных классов 0 (выше данный параметр назывался specifisity);
- *1-class score* — количество правильно предсказанных классов 1;
- *aucroc score* — площадь под roc кривой.

Для оценки каждого из четырех вышеперечисленных параметров здесь и далее мы будем использовать стратифицированный метод скользящего контроля по пяти блокам (*5-fold cross-validation*). Построим график зависимости качества классификации от количества соседей в методе KNN (см. Рис. 1).

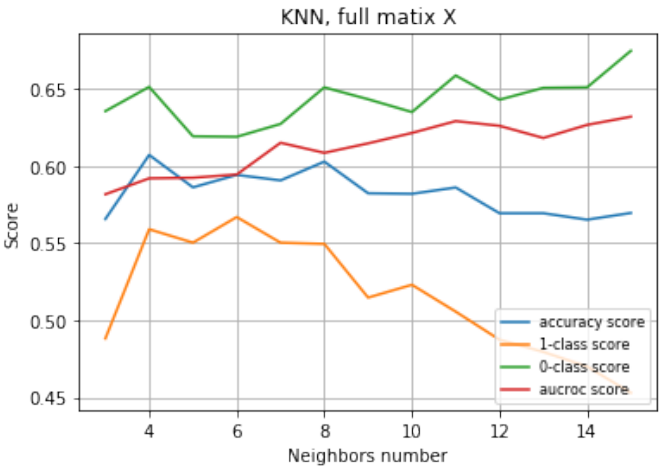


Рис. 1: График зависимости качества предсказания от количества соседей в методе ближайших соседей.

Из графика видно, что начиная с $N = 8$ характеристика *accuracy* начинает убывать. Максимальное значение параметра *accuracy* равно 60.7% (при $N = 4$). Максимум для *aucroc* равен 0.631 (при $N = 15$). С точки зрения

specifisity оптимальное количество соседей равно 15, максимальный процент угаданых пациентов, не откликнувшихся на лечение, равен 67.4%. Что касается предсказания класса 1, оптимальным параметром является $N = 6$, при котором качество предсказания достигает своего максимального значения 56.6%.

Далее протестируем метод опортных векторов (SVC). В качестве параметра будет выступать ядро классификатора. Рассмотрим сделующие ядра:

- линейное (linear);
- полиномиальное (poly) о степенью полинома, не более трех;
- радиальное (rbf);
- сигмоидальное (sigmoid).

Для каждого ядра оценим точность классификации (четыре характеристики, описанные выше) методом скользящего контроля и построим диаграмму (см. Рис. 2).

Как видно из графика, метод опорных векторов с радиальными сигмоидальным ядрами присваивает всем пациентам нулевой класс. Среди остальных двух ядер линейное ядро демонстрирует точность немного выше, чем полиномиальное.

Максимальное значение *accuracy* для классификатора SVC равно 59.8%, а для *aucroc* — 0.647.

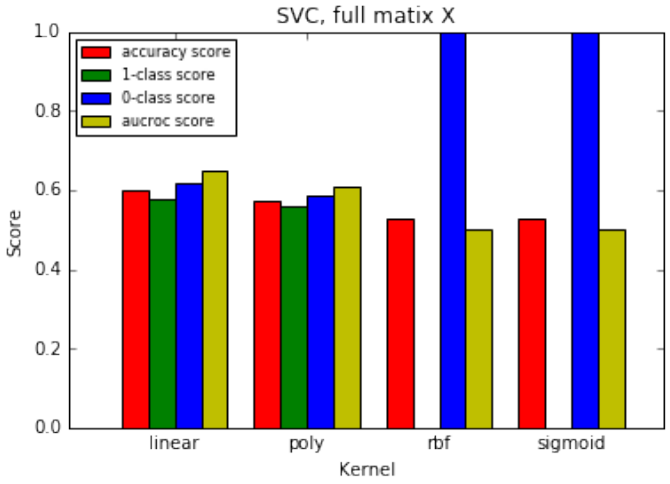
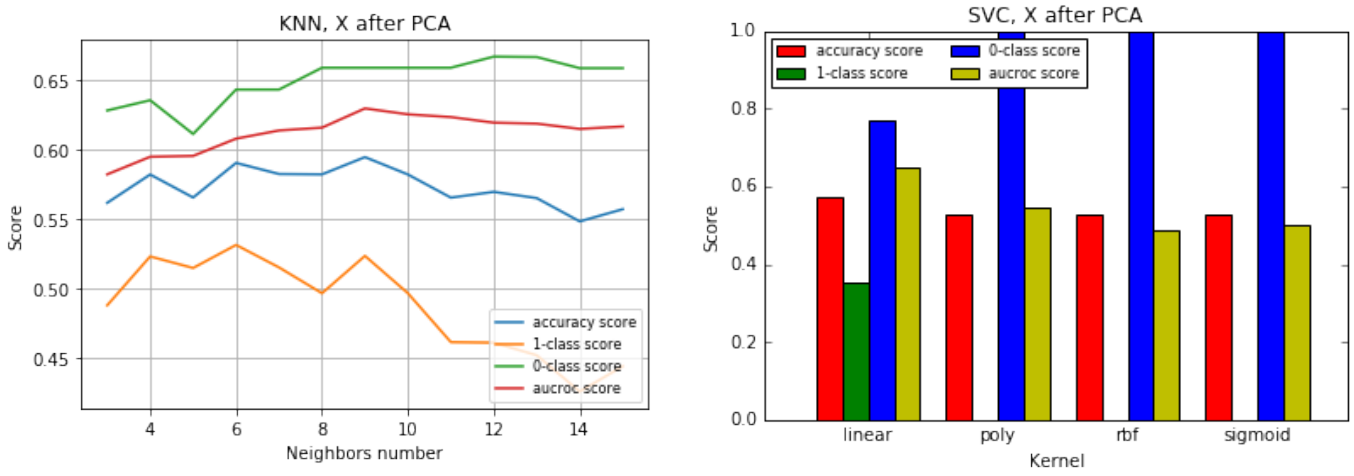


Рис. 2: График зависимости качества предсказания от ядра в методе опорных векторов.

4.2 Построение классификатора с применением метода главных компонент

Далее протестируем первый метод уменьшения размертости данных. А именно, применим метод главных компонент (PCA) и протестируем оба типа классификатора (KNN и SVC) на новой редуцированной матрице X_{PCA}. Как и прежде будем перебирать различные параметры классификаторов (для метода ближайших соседей - количество соседей, для метода опорных векторов - ядро). Построим



а) метод ближайших соседей

б) метод главных компонент

Рис. 3: Зависимость качества предсказания от параметров классификатора, построенного после применения метода главных компонент.

графики, аналогичные предыдущим, отображающие зависимость качества классификации от параметров классификаторов (см. Рис. 3).

Как видно из графиков, метод главных компонент практически не повлиял на качество классификации метода ближайших соседей (разве что, упала метрика *aucroc* при малых N). Существенные изменения произошли со вторым класификатором. А именно, метод опорных векторов с полиномиальным ядром стал приписывать всех пациентов одному и тому же 0 классу, как и в случае радиального и сигмоидального ядра. Качество классификации метода опорных векторов с использованием линейного ядра существенно возросло по сравнению с тем же классификатором, постоенным по полной матрице X, с точки зрения *specifisity*: процент правильно предсказанных представителей нулевого класса вырос на 15%. Площадь под roc-кривой практически не изменнилась. С другой стороны, значение оставшихся других характеристик уменьшилось: новое максимальное значение *accuracy* равно 57.3% (что меньше на 2.5% предыдущего показателя), а значение *1-class score* упало на 22.2%.

4.3 Метод экстремальной группировки

Метод экстермальной группировки был впервые изложен в 1970 году в статье Э.М.Бравермана "Методы экстремальной группировки параметров и задача выделения существенных факторов". Целью данного метода группировки является разбиение признаков на группы таким образом, чтобы внутри одной группы находились признаки, наиболее коррелирующие друг с другом, а признаки из разных групп коррелировали слабо. Количество групп, получаемых в результате разбиения, является параметром метода экстремальной группировки (далее общее коли-

чество признаком будем обозначать переменной n, а количество групп — переменной k). В своей статье Э.М.Браверман изложил несколько подходов к решению данной задачи, один из которых заключается в максимизации функционала

$$J(A_1, A_2, \ldots, A_k, f_1, f_2, \ldots, f_k) = \sum_{i \in A_1} (x_i, f_1)^2 + \sum_{i \in A_2} (x_i, f_2)^2 + \ldots + \sum_{i \in A_k} (x_i, f_k)^2.$$

Здесь:

- (x, y) обозначает корреляцию между случайными величинами x и y;
- случайные величины x_i $(i = 1, 2, \ldots, n)$ соответствуют признакам, подлежащим группировке;
- подмножества $A_i \subseteq \{1, 2, \ldots, n\}$ включают в себя индексы тех признаков, которые попали в i-тую группу $(i = 1, 2, \ldots, k)$;
- f_i $(i = 1, 2, \ldots, k)$ - случайные величины с единиичной дисперсией (далее будут называться центроидами).

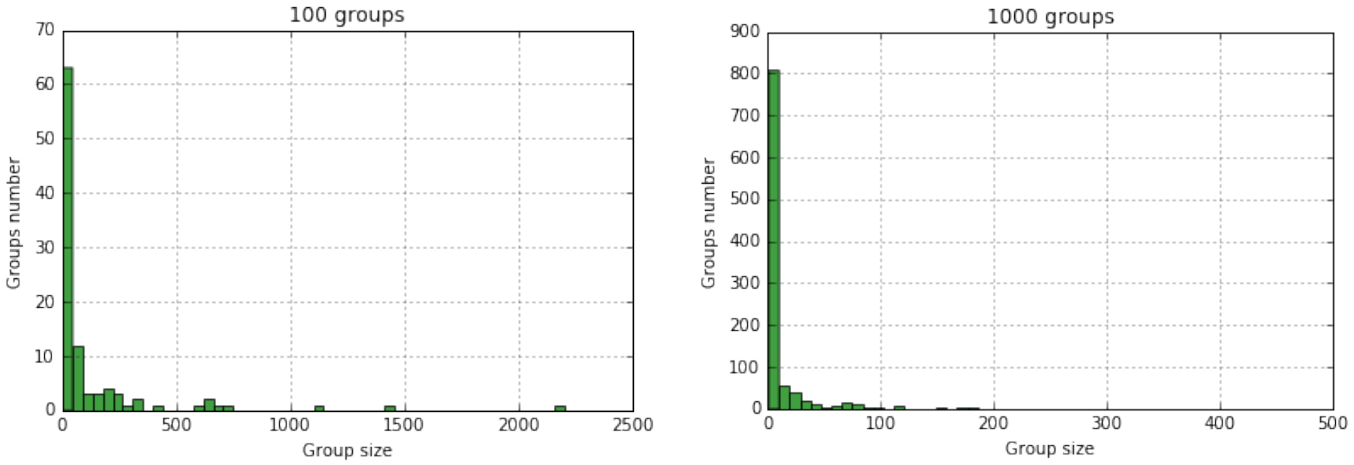
В рассматриваемом функционале подмножества A_i и центроиды f_i являются искомыми параметрами, по которым производится максимизация функционала, и подбираются итеративно до сходимости алгоритма.

Как уже было сказано выше, параметром метода экстремальной группировки, который задается пользователем, является количество групп k, на которые разбивается исходное множество факторов. Одной из интересных задач, возникающих в связи с применением данного метода, является анализ групп, получаемых в результате разбиения признаков, и исследование влияния параметра k на результат.

Применим метод экстремальной группировки к имеющимся данным, и построим разбиение факторов на 100-1000 групп с шагом 100. Одним из интересных вопросов является то, как меняется размеры построенных группы в зависимости от количества групп. Построим гистограммы численностей для случая 100 и 1000 групп (см. Рис. 4).

Как видно из графиков, большинство групп имеют сравнительно небольшой объем. В случае 100 групп более 60 из них состоит из не более, чем 50 факторов. Кроме того, среди ста полученнных групп три группы с наименьшим количеством факторов имеют объемы 1, 1, и 2 представителя. Три наибольших группы состоят из 1130, 1427, 2204 факторов соответственно.

В случае 1000 групп более 800 из них не превосходят 10 факторов. Более того, 380 из них состоят ровно из одного представителя. Три наибольших группы содержат 364, 394, 463 факторов соответственно.



а) разбиение на 100 групп

б) разбиение на 1000 групп

Рис. 4: Гистограммы численностей групп, полученных методом экстремальной группировки.

4.4 Построение классификатора на основе метода голосования

Следующий подход строит класификатор на основе групп, полученных методом экстремальной группировки, с применением метода голосования. Построение классификатора можно разбить на следующие этапы:

- в результате метода экстрмальной группировки получаются группы A_i $(i = 1, 2, \ldots, k)$;
- на основе обучающей выборки производится обучение k различных классификаторов $c_1, c_2, \ldots, c_k$, где i-тый классификатор строится на основе откликов y и признаков x_j, где $j \in A_i$ (то есть, тех факторов, чьи индексы попали в i-тую группу индексов);
- для тестовой выборки строится k предсказаний, каждое из которых получается применением соответствующего классификатора c_i к подгруппе факторов x_j, где $j \in A_i$.
- если среди полученных k предсказаний более, чем $\frac{k}{2}$ соответствуют классу 1, то результатом классификации будем считать класс 1, в противном случае ответом будем считать класс 0.

Для каждого разбиения на группы, полученного в предыдущем пункте, построим классификатор методом голосования. Для начала протестируем метод ближайших соседей. В качестве примера для случая $k = 100$ приведем график,

отображающий зависимость качества классификации от количества соседей (см. Рис 5).

Как видно из графика, *KNN* характеристики *aucroc* и *accuracy* практически не зависят от количества соседей в методе ближайших соседей. С другой стороны, как уже было упомянуто выше, характеристика *0-class score* характеризует долю правильных предсказаний для нулевого класса, таким образом, чем выше данная характеристика, тем меньшему количеству пациентов будет назначено неэффективное лечение. С точки зрения данной характеристики классификатор демонстрирует неплохие результаты (более 75%)

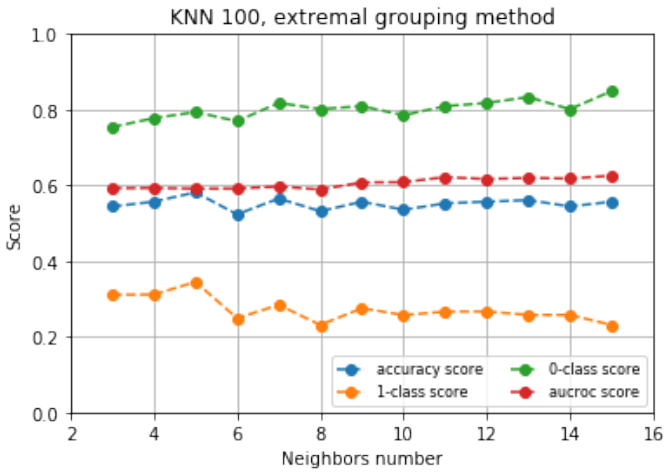


Рис. 5: Зависимость качества предсказания от количества соседей, метод голосования для 100 групп.

Наша конечная цель - получить максимальную точность предсказания (с точки зрения одной из четырех метрик), а также, получить оптимальные параметры классификатора, при котором достигается наилучшая точность. Поэтому для каждого разбиения на группы для каждой метрики качества вычислим максимальное (по N) значение точности предсказания, полученного методом голосования. Построим график зависимости максимального качества классификатора от количества групп в методе экстремальной группировки (см. Рис. 6).

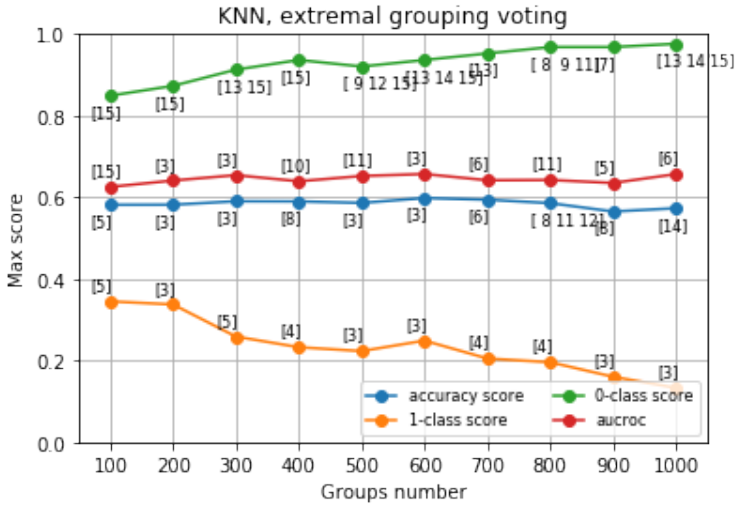


Рис. 6: Зависимость максимального качества предсказания от количества групп, метод голосования

Помимо максимальных значений метрик, на графике приведены массивы с наборами оптимальных параметров (в данном случае это оптимальное количество соседей), при котором для данного количества групп для данной метрики достигается ее максимальное значение. Так, например, если рассматривать количество групп, равным 100, то максимальнае значение метрики *0-class score* (или иначе *specifisity*) равно 0.83 и достигается, когда в методе количество соседей равно 15. Иными словами, при разбиении на 100 групп максимальное количество правильно предсказанных представителей 0 класса (83%) получается методом голосования, в котором каж-

дая группа "выносит вердикт"с помощью классификатора *KNN* с параметром $N = 15$.

Из графика видно, что для некторых характеристик качества и некоторых групп может быть несколько оптимальных парамтеров. Например, для 800 групп максимальное значение *accuracy* достигается, если в количество соседей равно 8, 11 или 12. Что интересно, для каждого из четырех параметров качества есть набор повторяющихся оптимальных параметров: для *accuracy* и *aucroc* наболее популярным выбором является 3 соседа, для *1-class* чаще всего в качестве оптимального количества соседей встречаются величины 3 и 4, а для *0-class score* — 13 и 15.

4.5 Анализ классификатора, построенного на основе всех центроидов

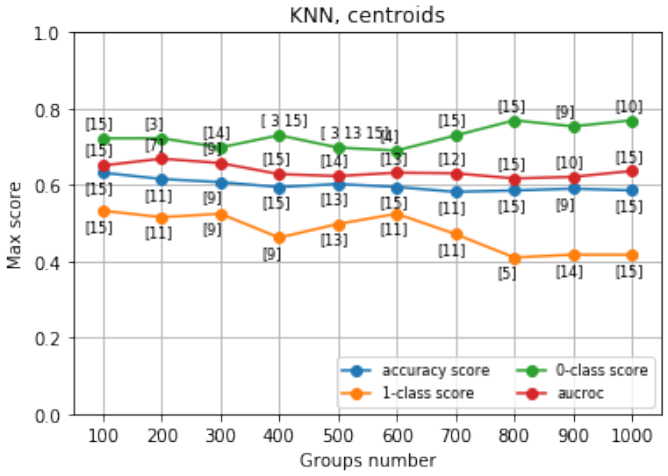


Рис. 7: Зависимость максимального качества предсказания от количества групп, метод центроидов

В следующей части исследования классификатор строился на основе центродов, полученных в методе экстремальной группировки, таким образом, количество признаков, используемых при построении классификатора равно количеству групп. Проделаем аналогичные эксперименты, как и в предыдущем пункте, и для каждого набора центроидов найдем максимальное значение метрик качества и оптимальные параметры. На графике видно, что предсказание, построенное с помощью центроидов, ничем не уступает предсказанию на основе метода голосования. Более того, с точки зрения всех метрик, кроме *specifisity*, мы видим некоторое существенное улучшение в работе классификатора. С другой стороны, количество оптимальных парамеров данного классификатора получилось чуть более разнообразное, чем в предыдущем пункте.

4.6 Поиск оптимальных центроидов

Одним из недостатков классификатора, построенного на основе центроидов, является то, что количество признаков равно количеству групп, на которые разбиваются все экспрессии. В случае, например, 1000 групп, размерность по-прежнему является сравнительно большой, относительно размера выборки (250). Таким образом, целью данного пункта является выделение некоторого "хорошего"подсемейства

центроидов. Рассмотрим три критерия выбора такого подсемейства.

(1) Рассмотрим произвольный центроид. Каждому пациенту будет соответствовать свое значение на оси центроида. Упорядочим пациентов на оси центроида и рассмотрим трех самых левых пациентов и трех самых правых пациентов на оси. Если одна из двух троек пациентов принадлежит одному классу, а не менее, чем двое пациентов из другой тройки принадлежат другому классу, то будем называть такой центроид "хорошим с точки зрения 6 крайних пациентов". Преимущество данных центроидов заключается в том, что их очень легко найти и что классификатор, предсказывающий отклик на основе данного центроида, гарантирует как минимум 2% правильных ответов.

(2) Рассмотрим произольный центроид. Построим классификатор на основе одного данного центроида. Если метрика *accuracy* превышает некоторого порогового значения, будем считать, что данный центроид "хороший с точки зрения построения однофакторного классификатора".

(3) Рассмотрим произвольный центроид. Посчитаем коэффициент корреляции с откликом на тарегную терапию (вектор Y). Если коэффициент корреляции превышает некоторое пороговое значение будем называть данный центроид "хорошим с точки зрения корреляции с откликом". Данные центроиды имеют "наибольшее влияние"на результат таргетной терапии, что означает, что их надо включать в список признаков в первую очередь.

Проведем подбор "хороших"центроидов для случая, когда количество групп равно 100. Получаем, что 18 из 100 рассматриваемых центроидов удовлетворяет критерию "хороший с точки зрения 6 крайних пациентов". Приведем номера групп, чьи центроиды оказались "хорошими":

$$0, 6, 21, 49, 57, 60, 64, 67, 68, 70, 72, 79, 81, 83, 84, 90, 91, 96.$$

В качестве эксперимента, вычислим количество "совсем хороших с точки зрения 6 крайних пациентов"центроидов: то есть таких центроидов, для которых одна из двух крайних троек полностью принадлежит одному классу, а вторая тройка полностью принадлежит другому классу. Из 18 "хороших"центроидов "совсем хороших"оказалось всего трое с номерами 57, 70 и 96.

Далее выделим список центроидов, которые являются "хорошими с точки зрения построения однофакторного классификатора". Проверка центроида на данный критерий осуществляется следующим образом. На первом шаге перебираем различные пороговые значения от 0 до 239 и считаем, что все пациенты, чьи порядковые номера на оси центроида меньше порогового значения принадлежат одному классу, а остальные пациенты, у кого порядковые номера больше порогового значения — принадлежат другому классу. Далее для каждого порогового

значения считаем ошибку классификации и минимизируем ошибку по пороговому значению.

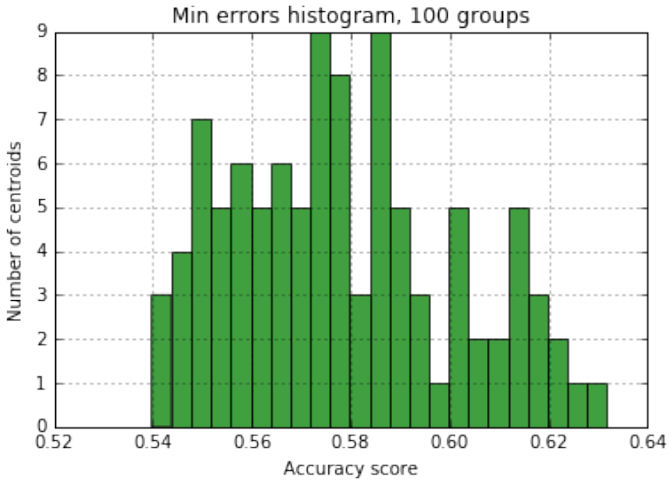


Рис. 8: Гистограмма распределения максимального качества предсказания с помощью одного центроида.

Для случая разбиения экспрессий на 100 групп, минимальное значение качества классификации, построенной по одному центроиду равно 54%, а максимальное - 63%. Построим гистограмму распределений для максимального качества предсказания с помощью одного центроида (см. Рис. 8).

Так как мы хотим уменьшить количество признаков, по которым строим классификатор, со 100 до не более, чем 25, то в качестве порогового значение выберем ошибку классификации, равную 0.6. Таким образом, в список "хороших с точки зрения однофакторной классификации"центроидов мы будем включать те, у которых минимальная ошибкла классификации не превышает 40%. Получаем следующий список из 21 центроида:

$$3, 6, 7, 8, 9, 10, 21, 25, 36, 38, 44, 50, 54, 67, 68, 70, 71, 74, 83, 89, 97.$$

Выделим группу хороших центроидов по признаку корреляции с откликом. Для имеющихся 100 центроидов вычислим их коэффициенты корреляции с откликом на лечение. Минимальное значение коэффициента корреляции(по абсолютному значению) равно 0.0023, максимальное - 0.24. Как и в предыдущем случае построим гистограмму распределения корреляций (см. Рис 9).

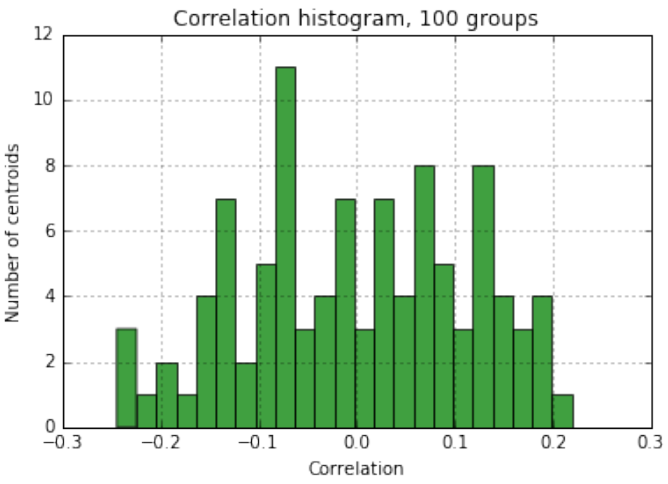


Рис. 9: Гистограмма распределения корреляции центроида с откликом Y.

Минимальное значение коэффициента корреляции(по абсолютному значению) равно 0.0023, максимальное - 0.24. Как и в предыдущем параграфе - построим гистограмму распределения.

Так как мы хотим уменьшить количество признаков до не более, чем 25, то в качестве порогового значения в данном случае возьмем величину 0.15 и будем считать "хорошими с точки зрения корреляции с откликом"такие центроиды, у

которых коэффициент корреляции по модулю превосходит 0.15. Таких центроидов получилось 20 штук:

$$0, 3, 6, 7, 8, 9, 10, 21, 25, 31, 36, 38, 44, 54, 59, 67, 68, 71, 74, 83$$

Для интереса, посмотрим, сколько "хороших с точки зрения 6 крайних пациентов"центроидов попало в список "хороших с точки зрения однофакторной классификации". Таких центроидов получилось 6 штук: 6, 21, 67, 68, 70, 83. Кроме того, посмотрим на пересечение "хороших с точки зрения 6 крайних пациентов"с "хороших с точки зрения корреляции с откликом": 6 элементов, номера групп соответствующих центроидов 0, 6, 21, 67, 68, 83. И, наконец, пересечение "хороших с точки зрения однофакторной классификации"и "хороших с точки зрения корреляции с откликом"состоит из 17 центроидов: 3, 6, 7, 8, 9, 10, 21, 25, 36, 38, 44, 54, 67, 68, 71, 74, 83. В пересечении всех трех методов лежат 5 центроидов: 6, 21, 67, 68, 83.

Для наглядности построим распределение пациентов на осях центроидов с данными номерами групп (см. Рис. **??**). На осях центроидов с номерами 6, 23, 67, 68 можно наблюдать длинные одноцветные промежутки. С точки зрения классификации, отсутствие большого количества подряд идущих пациентов из одного класса увеличивает "перемешанность"классов и, как следствие, уменьшает шансы на хорошую классификацию. Таким образом, график подтверждает, что выбранные центроиды могут быть эффективными с точки зрения использования их для построения предсказания.

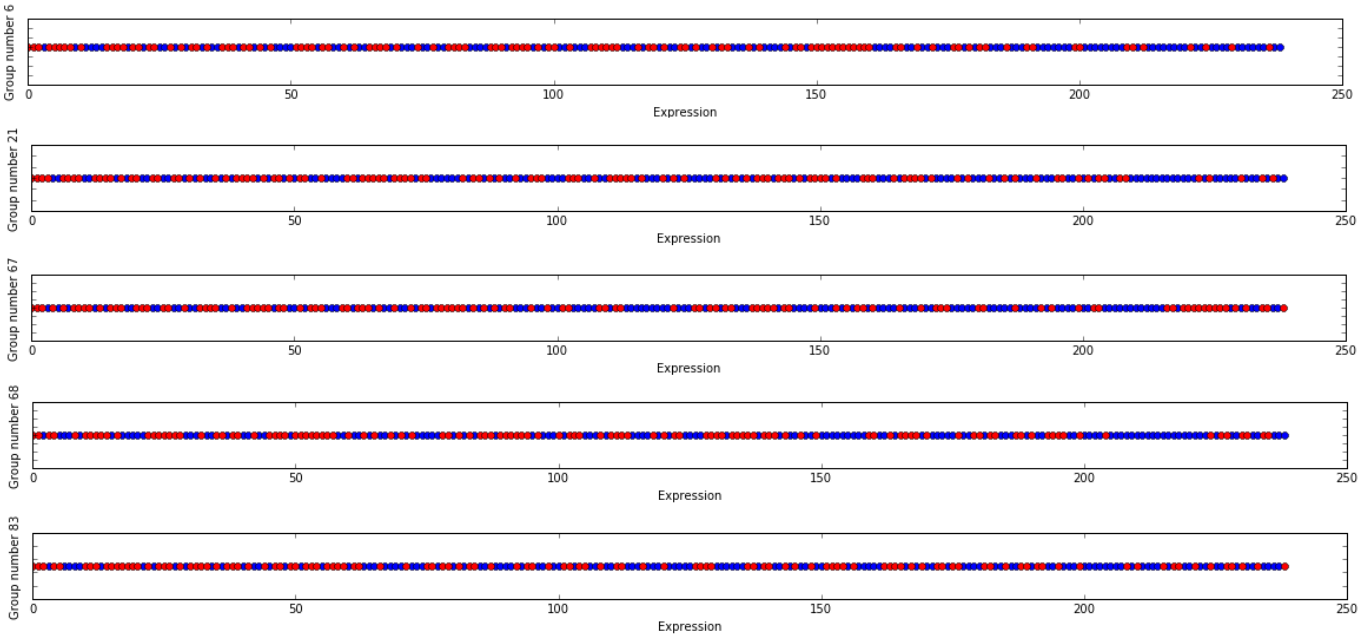


Рис. 10: Изображение пациентов на оси центроида.

Построим классификаторы на основе каждой из выделенных подгрупп центроидов и перечислим максимальные значения 4 характеристик качества предсказания. Получим следующую диаграмму, отражающую максимальное качество классификации в зависимости от метода выбора "хороших"центроидов (см. Рис 11). Стоит отметить, что все три метода выбора центроидов не сильно ухудшают результаты классификации по сравнению с классификацией, где в качестве призваков используются все центроиды. Таким образом, данный метод понижения размерности можно считать применимым для построения классификатора.

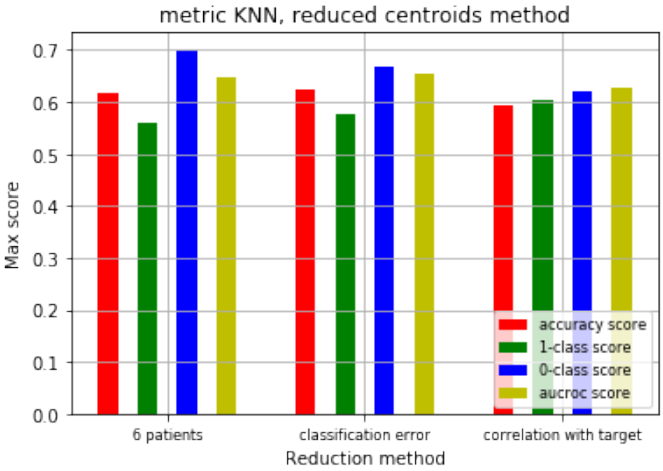


Рис. 11: Зависимость максимального качества классификации от метода выбора центроидов.