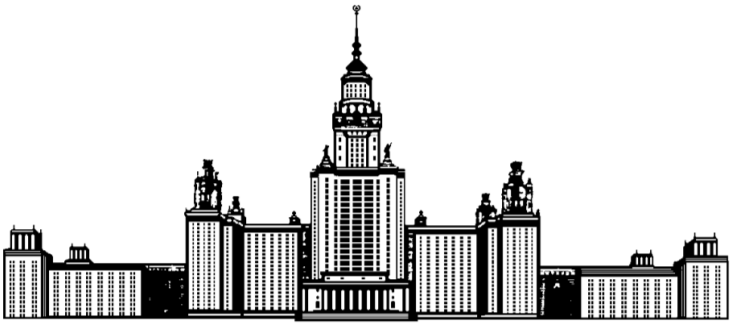
Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования



Корнилов Максим Алексеевич

Преобразование признакового пространства к матрице попарных расстояний в задачах на классификацию повышенного артериального давления

Курсовая работа

Научный руководитель:
д.ф.-м.н.
Сенько О. В.

Москва, 2020

Содержание

1 Введение

Рынок носимых устройств сильно расширился за последние 10 лет. За это время стали популярны Smart-часы и фитнесс-браслеты - устроства, одной из основных функций которых является отслеживание показателей здоровья носителя, что на данный момент ограничивается оцениванием пульса и давления. И если остлеживание пульса вышло на достаточно качественный и стабильный уровень, доступный устройствам всех ценовых категорий (а именно - погрешность в 10% у методик, основанных на фотоплетизмографии), то оценивание артериального давления всё ещё является особенностью устройств премиум-сегмента. При этом существующие на данный момент методики не могут достаточно стабильно определять давление пациента.

Поэтому может быть полезным разбить данную задачу на 2 подзадачи - сначала определить, повышенное или пониженное давление у человека, а затем уже - оценить значение, имея меньший разброс показателей.

В данной работе рассматривается альтернативный подход решения первой из этих задач путём преобразования признакового пространства к матрице попарных расстояний между объектами тренировочной выборки (которая в данном случае получается альтернативными методами), и построению аналогичной матрицы для данных, которые необходимо оценить.

2 Отбор признаков

Оценивание важности признаков и отбор наиболее важных среди них было решено проводить с помощью смешанных версий out-of-bag и permutation методов отбора признаков.

Базовые методы подразумевает оценивание важности параметра в методе Random Forest путём оценивания качества предсказаний лишь на тех деревьях, в подвыборку признаков которых не входил рассматриваемый признак (permutation), или на тех объектах, которые не входили в разбиение(out-of-bag). Главным недостатком подобного метода оказалась невозможность использовать данный метод вне ансамблевых методов, в построении которых используется bagging для получения подвыборок признаков.

В отличие от описанного выше алгоритма предлагается обобщить out-of-bag метод на любую модель машинного обучения. Суть данной идеи состоит в том, что вместо подбора деревьев, в которые не входит признак, тестируемый признак “портится” на всей модели. В зависимости от структуры признака, это может быть как простое перемешивание значений на разных объектах (любой тип признака), либо присваиванием всем объектам случайно сгенерированных значений той же сигнатуры, то есть имеющим такие же среднее и стандартное

отклонение (численные признаки).

Критерием полезности признака в данном случае будет повышение или понижение качества работы модели. так как текущая задача - задача классификации, то метрикой может служить roc-auc-score.

Для большей точности оценки важности параметра можно учитывать не только сам факт повышения/понижения качества работы модели, но и значение данного повышения или понижения. то есть вес данного признака рассчитывается, как

$$\delta_i = max(0, \sum_{k=1}^{n}(S - S_{ik})),$$

где δ_i - первичный вес $i-$го признака, S - результат работы данного метода на наборе данных без "порчи" переменных, а S_{ik} - результат работы метода с "испорченым" i-ым признаком на k-ом запуске

В дальнейшем рассмотрении учитываются только те значения δ_i, которые удовлетворяют порогу $\delta_i > 0.05$ - оно необходимо для того, чтобы не получались невероятно большие отрицательные значения. Следующим шагом от каждого оставленного изначального веса берется логарифм и находится глобальный минимум среди всех полученных значений.

$$\psi_i = ln(delta_i)$$
$$min_\psi = min(\psi_i | i : \delta_i > 0.05$$

Итоговый вес каждого признака получается прибавлением к каждому из ψ_i значения их глобального минимума, тем самым получая неотрицательные веса признаков:

$$\omega_i = \psi_i + min_\psi$$

Применив данный метод к тренировочной задаче, используя CatBoost как базовую модель, можно получить следующие результаты

На рисунке 1 можно увидеть, что среди общего числа признаков (87) можно выделить треть наиболее полезных, среди которых около 10 имеют наибольшую значимость, и значимость остальных признаков значительно меньше.

Для оценки качества работы данного метода отбора его можно сравнить с методами оценивания важности параметров, качество которых уже было доказано: В качестве сравнения оценки важности параметров был взят встроенный в алгоритм градиентного бустинга CatBoost метод

На рисунке 2 видно, что данные модели одинаково определили набор ключевых (самых значимых 20) признаков, дающих наибольший вклад в суммарный вес и дали им схожий порядок и схожие показатели. Так что построенный в

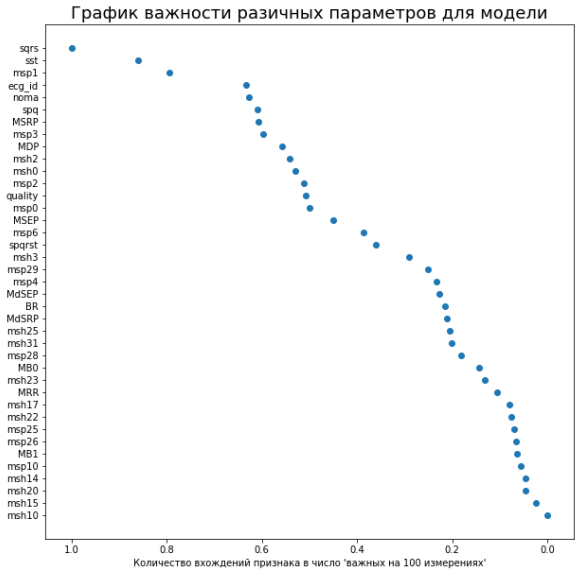


Рис. 1: оценивание важности признаков на тестовой задаче

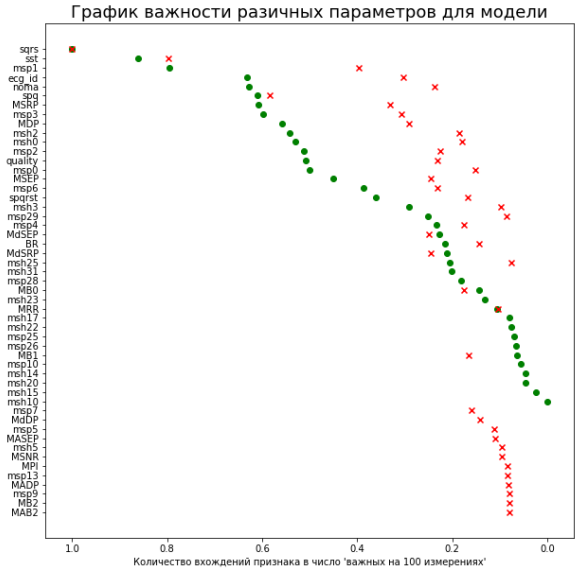


Рис. 2: сравнение методов оценки важности признаков, X - метод, встроенный в модель CatBoost

данном пункте алгоритм считаю возможным рассматривать, как метод оценки важности признаков.

В итоге данных действий был получен поднабор признаков, наиболее по-

лезных среди полного набора и их веса, на основе которых будут проведены дальнейшие преобразования.

3 Построение матрицы попарных расстояний

Для тренировочного набора данных матрица попарных расстояний строится следующим образом: Пусть $A = (a_{ij}) \in R^{n*l}$ - массив тренировочного набора данных построчно. Точка a_{ij} - j-ая переменная i-го объекта.

Рассмотрим получение матрицы попарных расстояний для каждого из признаков в отдельности. Пусть $B^i = b^i_{km} \in R^{n*n}$ - матрица попарных расстояний между объектами тренировочной выборки для i-ой переменной. Рассмотрим значение каждой точки этой матрицы в отдельности:

$$b^i_{km} = |a_{ki} - a_{mi}|$$

В дальнейшем каждая из матриц B^j нормализуется - приводится к набору значений $[0, 1]$, что достигается путём деления матрицы на значение её максимального элемента.

В итоге данного действия получается набор $B^i, i \in [1, l]$ матриц, соответствующих попарному расстоянию между объектами тренировочной выборки по i-му признаку. Матрица попарных расстояний между объектами тренировочной выборки получается суммированием данного набора матриц:

$$D_{train} = \sum_{i=1}^{l} B^i$$

Для предсказываемых данных матрица попарных расстояний строится аналогичным образом: Пусть $A = a_{ij} \in R^{n*l}$ - матрица тренировочного набора данных, $C = (c_{ij} \in R^{m*l}$ - матрица предсказываемого набора данных. Для каждого из признаков аналогичным описываемым выше образом построим матрицу попарных расстояний по этому признаку:

$B^i = (b^i_{kp}) \in R^{m*n}$, для $i \in [1, l]$ поэлементно строится следующим образом:

$$b^i_{kp} = |c_{ki} - a_{pi}|$$

После приведения матриц B^i к формату значений $[0, 1]$ получим матрицу попарных расстояний взвешенным суммированием:

$$D_{val} = \sum_{i=l}^{l} \omega_i * B^i$$

Итогом данных преобразований получаются 2 матрицы попарных расстояний для тренировочного и валидационного набора данных, к строкам которых подходят те же метки, что и к строкам исходных данных.

Для упрощения рассчетов можно привести элементы матрицы тренировочных данных к промежутку значений $[0, 1]$, и валидационных к промежутку $[0, \frac{max(D_{val})}{max(D_{train}}]$, разделив все элементы обеих матриц на $max(D_{train})$.

4 Сравнение качества работы на CatBoost

В качестве одной из моделей, на которой было проведено сравнение качества работы измененного признакового пространства относительно исходных данных был взят алгоритм градиентного бустинга, основанный на решающих деревьях - CatBoost. Это алгоритм, разработанный инженерами Yandex, используемый в поиске, рекомендательных системах, "Алисе"и во многих других проектах. Он показал хороший результат относительно других популярных моделей градиентного бустинга (а именно, LightGBM, HGBoost и H2O). Особенностью данной модели является уникальный алгоритм, нацеленный на ускорение схождения модели, при этом оставляя высокое качество работы.

Мною были протестированы 4 набора данных по разделению давления пациентов на повышенное и пониженное. На рисунке №3 можно увидеть, что в зависимости от набора данных лучшее качество работы варьируется между изначальным набором данных и преобразованием к матрице попарных расстояний. Если усреднить результаты по всем 4 наборам данных, то усредненный лучший результат среди методов, использующих исходные признаки будет 0.85, а на матрицах попарных расстояний - 0.84.

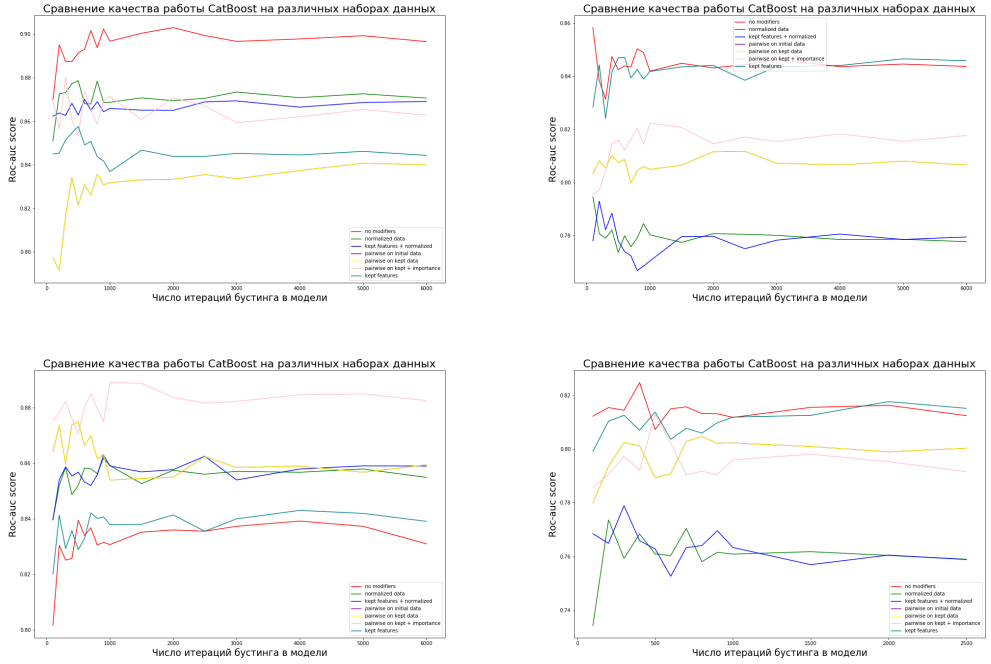


Рис. 3: Сравнение качества работы алгоритмов подбора признаков на CatBoost

Наилучшим из методов, использующих матрицу попарных расстояний оказался тот, что использует подвыборку "полезных"признаков (отобранную по порогу от максимального значения), и учитывает в построении матрицы попарных расстояний веса данных признаков.

5 Сравнение качества работы на LightGBM

Кроме CatBoost проверим работу матрицы попарных расстояний на другом алгоритме градиентного бустинга - LightGBM. В отличие от CatBoost, он ориентирован на ускорение схождения алгоритма без последующей оптимизации (достаточно быстро модель попадает в локальный минимум, что останавливает изменение качества работы).

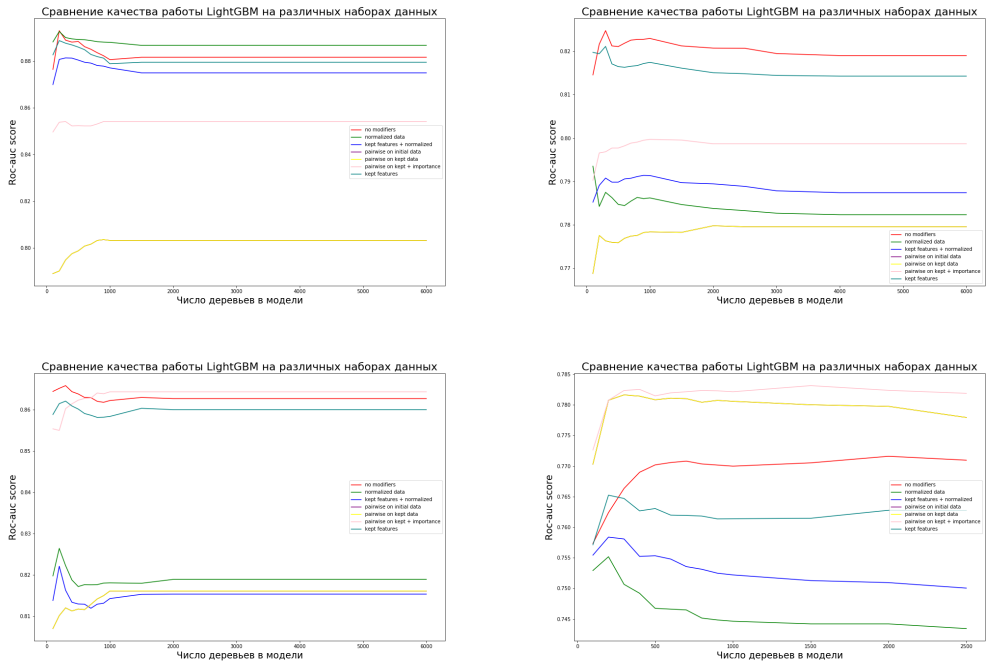


Рис. 4: Сравнение качества работы алгоритмов подбора признаков на LightGBM

Эксперименты с данной моделью подтверждают выводы, полученные при использовании CatBoost. Перевод к попарным расстояниям так же дает схожие результаты с классической моделью (в данном случае - 0.8325 в среднем у классических методов против 0.8275 у матриц попарных расстояний).

Среди моделей, использующих матрицы попарных расстояний лучший результат так же был показан моделью, использующей фильтрацию признаков + учитывающей важность оставшихся признаков.

6 Выводы

В данной работе был описан метод построения матрицы попарных расстояний для данных пациентов с целью упрощения решения задач классификации повышенного артериального давления.

Для ускорения предсказываний и повышения качества работы данного метода были проведены следующие эксперименты, преобразовывающие данные:

- была проведена оценка полезности исходных признаков (смесь permutation и out-of-bag методов оценивания полезности признаков)
- по данной оценке полезности был произведен отсев признаков (были оставлены 30% наиболее полезных).
- среди самих данных так же была произведена оценка (для матрицы попарных расстояний) - были удалены выбросы из данных

В результате данных экспериментов был получен алгоритм генерации нормированной по признакам матрицы попарных расстояний между объектами тренировочного набора данных и метод преобразования тестовых данных, опирающийся на тренировочные. Так же был построен метод оценки важности параметров, основанный на величине изменения качества работы алгоритма при "порче"данных в параметре. На тестовых наборах данных этот алгоритм показал схожие результьтаты с методами, используемыми для алгоритма решающего леса.

Для сравнения качества работы классификации на преобразованных признаках и исходных были применены алгоритмы градиентного бустинга, которые показывали стабильно хорошие результаты на задачах подобного типа.

При тестировании на алгоритмах CatBoost и LightGBM градиентого бустинга был получен хороший результат, схожий с показателями стандартных методов - на алгоритме CatBoost было достигнуто среднее качество в 0.84 roc-auc и 0.83 в случае LightGBM.

В результате данных тестов у алгоритма были выделены следующие преймущества:

- уменьшается необходимость преобразовывания входных данных
- упрощается представление данных в памяти, так как они все переходят в численный формат

Но так же у него можно выделить следующие недостатки:

- для предсказания значения необходимо держать в памяти исходный тренировочный набор данных, который используется для преобразования признаков

- в зависимости от построения попарных расстояний теряется возможность оценивать важность исходных признаков

- появляется нулевая диагональ матрицы попарных расстояний на тренировочном наборе. В качестве исправления было решено заменять её либо средним, либо медианным, либо максимальным среди значений.

- при большом размере тренировочной выборки значительно увеличивается размер признакого пространства (так, например, в представленной задаче происходит переход от 90 признаков к 450-850), в связи с чем в моделях, не способных на "обнуление"признаков наблюдается значительное повышение времени работы.

Данный метод лучше подходит для задач с небольшим объёмом тренировочной выборки, так как в обратном случае получается слишком объёмное признаковое пространство (что вкупе с необходимостью держать исходные данные в памяти для преобразования тестовых данных выливается в большие затраты по памяти). Преймуществом этого метода для малых выборок также в том, что вне зависимости от объёма тренировочного набора данных он показывает стабильный результат (не замечено большого падения качества работы при уменьшении выборок в 5, 10 раз).

Список литературы

[1] Phyu, Thu Oo, Nyein. (2016). Performance Comparison of Feature Selection Methods. MATEC Web of Conferences. 42. 06002. 10.1051/matecconf/20164206002.

[2] Баранов А.А., Намазова-Баранова Л.С., Смирнов И.В., Девяткин Д.А., Шелманов А.О., Вишнёва Е.А., Антонова Е.В., Смирнов В.И. Технологии комплексного интеллектуального анализа клинических данных // Вестник РАМН. 2016. №2.

[3] catboost. URL: https://catboost.ai (Дата обращения: 28.04.2020)

[4] scikit-learn. URL: https://scikit-learn.org/stable/index.html (Дата обращения: 27.04.2020)

[5] SciPy. URL: https://docs.scipy.org/doc/scipy/reference/index.html (Дата обращения: 27.04.2020)

[6] LightGBM. URL: https://lightgbm.readthedocs.io/en/latest/ (Дата обращения: 28.04.2020)

[7] Hastie Trevor, Tibshirani Robert, Friedman Jerome The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2009

[8] Bishop C. M. Pattern Recognition and Machine Learning. — Springer, 2006.