Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Серов Сергей Сергеевич

# Разработка методов оптимизации признакового пространства

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

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

Москва, 2018

# Содержание

**Аннотация**

Одной из важных задач машинного обучения является поиск оптимальных признаковых подпространств в данных высокой размерности. Существует множество методов решения этой задачи, но есть работы, в которых показано снижение эффективности некоторых алгоритмов отбора с увеличением размерности признакового пространства. В связи с этим необходимость продолжения исследований в этой области очевидна. В работе предлагается два метода, основанных на обучении наборов одномерных базовых предикторов, соответствующих отдельным признакам. Далее производится выбор подмножества предикторов, удовлетворяющего критериям максимальности точности прогноза и наибольшей разнородности. Полученные в работе результаты свидетельствуют о том, что предложенные методы показывают себя лучше существующих подходов на некоторых прикладных задачах высокой размерности.

# 1 Введение

В настоящее время машинное обучение активно применяется во многих сферах человеческой деятельности. В частности, в последние годы методы машинного обучения используются в решении медицинских задач и задач анализа сигналов. В этих областях науки часто возникают задачи, в которых в качестве обучающих выборок используются данные очень высокой размерности. Например, таковыми являются данные о генотипе человека, а также данные всевозможных медицинских исследований, которые часто отражают информацию, собранную в реальном времени с некоторой частотой. В задачах анализа сигналов признаковые пространства высокой размерности могут получаться, например, вследствие искусственной генерации новых признаков при поиске закономерностей. В связи с этим возникают проблемы, связанные с тем, что далеко не все признаки являются информативными и обучение моделей сильно усложняется для большой размерности признакового пространства. Таким образом, актуальной становится задача поиска оптимального признакового подпространства, которое содержало бы только информативные признаки, и при этом обучение на сокращенных данных не приводило бы к ухудшению качества прогноза.

Существует достаточно много методов решения этой задачи, и часто используется следующая классификация. Выделяют три типа методов:

- фильтрационные,
- обёрточные,
- встроенные.

Фильтрационные алгоритмы основаны на ранжировании признаков в соответствии с их индивидуальной информативностью, обёрточные — на оценке прогностической способности модели на различных подмножествах признаков, а встроенные — являются неотъемлемой частью процесса обучения основной модели.

Одними из самых популярных алгоритмов являются отбор с помощью регуляризации [8], отбор с помощью решающих лесов [4] и методы отбора, основанные на оценке информативности признаков (например, на взаимной информации между признаками и целевой переменной) [10].

Высокую эффективность в настоящее время показывают методы отбора, использующие регуляризацию, однако теоретических результатов, гарантирующих достижение максимально возможного качества отбора, не было получено. Более того, существуют работы [2] [1] [5], в которых было показано, что эффективность таких методов может значительно снижаться при увеличении размерности признакового пространства до очень больших и сверхбольших значений (более, чем $10^4 - 10^5$ признаков). В связи с этим продолжение работы в области создания методов отбора признаков для задач с данными очень высокой размерности актуально.

Цель данной работы заключается в разработке новых методов решения задачи оптимизации признакового пространства, которые не будут иметь ограничений на дискретность природы признаков и будут показывать высокую эффективность в задачах с высокой размерностью признакового пространства.

В работе сначала обсуждаются существующие методы решения данной задачи и разбирается алгоритм FAST [7], затем дается подробное описание предлагаемых методов и приводятся результаты экспериментов, поставленных на реальных данных из задач предсказания роста человека по информации о его генах и предсказания систолического давления по данным ЭКГ и фотоплетизмограммы.

# 2 Постановка задачи

## 2.1 Определения и обозначения

Рассмотрим стандартную задачу машинного обучения. Пусть дана обучающая выборка $(\mathbb{X}_{train}, \mathbb{Y}_{train})$, где $\mathbb{X}_{train} \in \mathbb{R}^{N \times M}$ — матрица признаковых описаний объектов, а $\mathbb{Y}_{train} \in \mathbb{R}^N$ — вектор значений целевой переменной (правильных ответов) для описанных объектов. Она содержит информацию о некоторых объектах из генеральной совокупности $\Omega = (\mathbb{X}, \mathbb{Y})$. Требуется, основываясь на информации об объектах из обучающей выборки, построить модель

$$G : \mathbb{R}^M \to \mathbb{R}$$

такую, чтобы она была оптимальной по заданному критерию качества

$$S : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

в смысле математического ожидания по всевозможным объектам из генеральной совокупности. Формально эта задача записывается так:

$$\mathbb{E}_\Omega S(G(X), Y) \to \max_G . \tag{1}$$

На практике при решении задачи чаще всего доступ ко всей генеральной совокупности отсутствует, поэтому для независимой оценки качества моделей используется валидационная подвыборка $(\mathbb{X}_{val}, \mathbb{Y}_{val})$ генеральной совокупности, которая должна быть репрезентативной, то есть достаточно полно и точно отражать зависимости, присутствующие в генеральной совокупности.

В некоторых задачах размерность $M$ пространства признаков может быть очень большой ($> 10^4 - 10^5$). Например, такие ситуации часто возникают в биологических задачах, использующих данные о генотипе живых существ, а также в задачах анализа сигналов. В связи с этим начинают проявляться следующие проблемы:

- наличие слабо коррелирующих или не коррелирующих с ответом признаков;

- наличие очень похожих или дублирующих друг друга признаков;
- высокая сложность вычислений для высокой размерности признакового пространства.

## 2.2 Задача оптимизации признакового пространства

В связи с этим становится актуальной задача оптимизации признакового пространства, решение которой преследует цель избавиться от вышеуказанных проблем. Запишем эту задачу более формально. Введем множество номеров признаков $F$ и будем обозначать за $\{f\}$ некоторое его подмножество.

Тогда новые матрицы, содержащие информацию о признаках с номерами $\{f\}$ объектов $X$, будем обозначать за $X_{\{f\}}$.

Наконец, приходим к более общей постановке задачи машинного обучения, включающей оптимизацию признакового пространства:

$$\mathbb{E}_\Omega S(G(X_{\{f\}}), Y) \to \max_{G,\{f\}}. \tag{2}$$

Здесь при записи $G(X_{\{f\}})$ имеется в виду, что функция $G$ принимает на вход вектор размерности $M$, но при работе учитывает информацию только о признаках с номерами из множества $\{f\}$.

Одновременная оптимизация по модели $G$ и подмножеству номеров признаков $\{f\}$ обычно является сложной, поэтому часто используется последовательная оптимизация сначала по $\{f\}$, а затем — по $G$. Далее в этой работе нас будет интересовать только оптимизация по подмножеству номеров признаков $\{f\}$ при фиксированных параметрах модели $G$. Поэтому основная задача, которая решается далее, принимает вид:

$$\mathbb{E}_\Omega S(G(X_{\{f\}}), Y) \to \max_{\{f\}}. \tag{3}$$

В терминах функции потерь

$$L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

эта задача может быть переписана в виде:

$$\mathbb{E}_\Omega L(G(X_{\{f\}}), Y) \to \min_{\{f\}}. \tag{4}$$

Везде далее под выражениями вида $G(\mathbb{X})$ будем понимать последовательное применение модели $G$ ко всем объектам из выборки $\mathbb{X}$ и объединение результатов в вектор.

# 3 Существующие методы решения

Глобальный оптимум задачи оптимизации (3) может быть найден без дополнительных предположений только методом полного перебора всех подмножеств множества $F$, сложность которого равна $O(2^M)$, то есть является экспоненциальной, и который при достаточно большом $M$ не может быть произведен за разумное время.

Поэтому на практике чаще всего переходят к поиску локальных оптимумов задачи (3), который уже может быть произведен за полиномиальное время, то есть имеет сложность $O(M^\alpha), \alpha \geq 0$.

Известным методом, например, является отбор признаков с помощью $L_1$-регуляризации (так называемой LASSO-регуляризации) [8]. В статье [7] 2013-го года также предлагается алгоритм FAST отбора признаков, основанный на их кластеризации.

Далее подробнее опишем эти методы.

## 3.1 Отбор признаков с помощью LASSO-регуляризации

В 1996-м году группой канадских ученых был предложен следующий метод регуляризации, который, как будет показано ниже, может быть использован для отбора признаков.

Пусть модель $G(X)$, решающая задачу машинного обучения, содержит в себе вектор параметров $\omega \in \mathbb{R}^M$, каждый из которых отвечает за важность признака с соответствующим номером при решении задачи. Необходимо минимизировать среднюю ошибку модели на генеральной совокупности:

$$\mathbb{E}_\Omega L(G(X), Y) = -\mathbb{E}_\Omega S(G(X), Y) \to \min_\omega . \tag{5}$$

При вычислении оптимальных значений параметров $\omega$ в процессе обучения на обучающей выборке важно избежать переобучения, то есть подбора слишком точных значений параметров для минимизации функции ошибки на объектах обучающей выборки. Для этого авторы предлагают модифицировать функцию ошибки следующим образом:

$$L(G(X), Y) = L(G(X), Y) + \lambda \sum_{i=1}^{M} |\omega_i|. \tag{6}$$

Таким образом, дополнительный член в функции ошибки не позволяет параметрам $\omega$ принимать большие по модулю значения, что значительно снижает эффект переобучения. Однако этот метод обладает еще одним полезным свойством: при достаточно большом значении коэффициента $\lambda$ оптимальная точка такой функции ошибки будет разреженной по параметрам $\omega$, то есть найдутся нулевые компоненты этого вектора.

Ниже приведем авторскую иллюстрацию, демонстрирующую разреживающее свойство LASSO-регуляризации в сравнении с $L_2$-регуляризацией (в этом случае регуляризационный член имеет вид $\lambda \sum_{i=1}^{M} \omega_i^2$). Эллипсами обозначены линии уровня функции потерь, а в начале координат расположены фигуры, задающие ограничения, отвечающие различным регуляризационным членам.

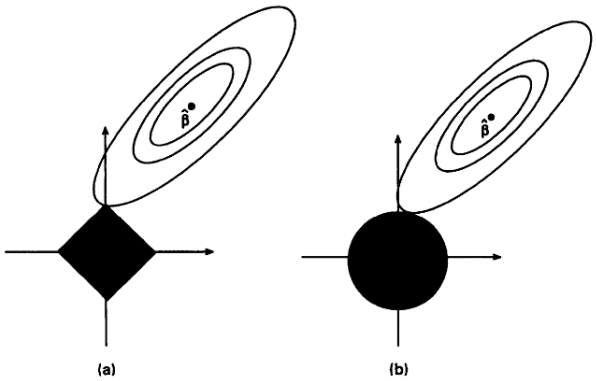


Рис. 1: Сравнение оптимальных точек функции ошибки с разными регуляризационными членами. (a) — LASSO-регуляризация, (b) — $L_2$-регуляризация.

Это полезное свойство может быть использовано для отбора признаков, если, например, каждая компонента вектора $\omega$ отвечает одному из столбцов матрицы признакового описания объектов. Так, в модели линейной регрессии LASSO-регуляризация может быть применена напрямую. После обучения модели с модифицированной таким образом функцией потерь часть признаков объектов генеральной совокупности не будет использована при построении окончательного прогноза.

## 3.2 Алгоритм FAST

Алгоритм FAST [7], предложенный в 2013-м году группой китайских ученых, является алгоритмом оптимизации признакового пространства, основывающимся на кластеризации (разбиении признаков на группы) и взаимной информации [9]. Он применим только для признаковых описаний объектов, в которых все признаки дискретны, то есть могут принимать конечное число значений. Далее кратко приведем идею этого метода.

Пусть $p(x)$ — априорные вероятности для всех возможных значений случайной величины $x$. Тогда энтропию этой случайной величины можно рассчитать по следующей формуле:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{7}$$

Пусть теперь $p(x|y)$ — апостериорные вероятности для всех возможных значений случайной величины $x$ при известном значении $y$. Тогда условная энтропия $x$ при условии $y$ может быть вычислена по следующей формуле:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{8}$$

Взаимной информацией называется величина

$$MI(X|Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{9}$$

имеющая смысл количества информации, которую несет одна случайная величина о другой, и являющаяся нелинейным аналогом корреляции.

Далее авторы статьи вводят понятие симметрической неопределенности следующим образом:

$$SU(X,Y) = \frac{2MI(X|Y)}{H(X)+H(Y)}. \tag{10}$$

После введения обозначений алгоритм выглядит так:

1. Для всех имеющихся признаков вычислить их симметрическую неопределенность с целевой переменной: $SU(X_i, Y)$. Удалить из рассмотрения признаки, для которых вычисленное значение ниже заданного порога.

2. Для всех оставшихся признаков вычислить их попарную симметрическую неопределенность $SU(X_i, X_j)$. Создать полный граф $H$, где каждой вершине соответствует один из оставшихся признаков $X_i$. Каждому ребру между вершинами $(i, j)$ присвоить вес, равный вычисленному значению $SU(X_i, X_j)$.

3. Построить минимальное остовное дерево графа $H$ по алгоритму Прима [6].

4. Получить множество деревьев (лес) $\{T_i\}_{i=1}^{K}$ с помощью удаления ребер из остовного дерева по следующему правилу: удалить ребро $(i, j)$, если выполнено условие:

$$SU(X_i, X_j) < SU(X_i, Y) \wedge SU(X_i, X_j) < SU(X_j, Y). \tag{11}$$

5. Составить итоговое подмножество признаков следующим образом. Из каждого дерева $T_i$ полученного леса выбрать один признак с тем номером, которому соответствует наибольшее значение симметрической неопределенности между признаками из этого дерева и целевой переменной:

$$\{f\} = \{j_i\}_{i=1}^{K},$$

$$j_i = \arg\max_{j \in T_i} SU(X_j, Y), \forall i \in 1, \ldots, K. \tag{12}$$

Описанный метод является эффективным по вычислительной сложности ($O(M * \log^2 M)$) и показывает хорошие результаты в задачах с высокой размерностью признакового пространства. Однако он существенно основывается на предположении, что все признаки в признаковых описаниях объектов, а также целевые переменные могут принимать лишь конечное число значений. Только в этом случае информационную энтропию и взаимную информацию для признаков можно определить. Ниже нами будет предложена модификация алгоритма FAST, позволяющая использовать его также для вещественных признаков.

# 4 Предлагаемые методы решения

Далее опишем предлагаемые нами методы решения задачи оптимизации (3).

Введем необходимые понятия и обозначения.

Пусть по каждому признаку $f_i \in F$ построен предиктор $G_i(X_i)$ для предсказания значений целевой переменной. Выпуклой комбинацией предикторов назовем предиктор $G(X)$ такой, что:

$$G(X) = \sum_{i \in \{f\}} c_i G_i(X_i), \tag{13}$$

где

$$\sum_{i \in \{f\}} c_i = 1, c_i \geq 0, \forall i \in \{f\}.$$

Поскольку в решаемой задаче (3) мы не имеем данных об априорной важности тех или иных признаков из множества $F$, то есть основания положить все коэффициенты $c_i$ равными $\frac{1}{|\{f\}|}$. Таким образом:

$$G(X) = \frac{1}{|\{f\}|} \sum_{i \in \{f\}} G_i(X_i). \tag{14}$$

Для каждого предиктора $G_i(X_i)$ мы можем вычислить векторы предсказаний целевой переменной для всех объектов обучающей выборки — $G_i(\mathbb{X}_{train,i})$, а также расстояния $\rho(G_i(\mathbb{X}_{train,i}), Y)$ в пространстве прогнозов от каждого предиктора до вектора истинных значений целевой переменной и попарные расстояния $\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j}))$ между ответами предикторов в пространстве прогнозов. Здесь расстояние $\rho(G_i(\mathbb{X}_i), \mathbb{Y})$ задается в соответствии с поставленной задачей, например как $\frac{1}{|\mathbb{X}_{train}|} \sum_{(X,Y) \in (\mathbb{X}_{train}, \mathbb{Y}_{train})} L(G_i(X_i), Y)$ или как стандартное евклидово расстояние.

## 4.1 Отбор с использованием функционала ошибки выпуклой комбинации

Введем функционал ошибки выпуклой комбинации предикторов:

$$\mathcal{L}(G, X, Y, \{f\}) = \frac{1}{|\{f\}|} \sum_{i \in \{f\}} \rho(G_i(X_i), Y) - \frac{1}{2|\{f\}|^2} \sum_{(i,j) \in \{f\} \times \{f\}} \rho(G_i(X_i), G_j(X_j)). \tag{15}$$

Для решения задачи (3) перейдем к задаче

$$\frac{1}{|\mathbb{X}_{val}|} \sum_{(X,Y) \in (\mathbb{X}_{val}, \mathbb{Y}_{val})} \mathcal{L}(G, X, Y, \{f\}) \to \min_{\{f\}}, \tag{16}$$

где под $|\mathbb{X}_{val}|$ здесь и далее понимается количество объектов в выборке $\mathbb{X}_{val}$.

Ее смысл в том, что выполняется поиск подмножества номеров признаков $\{f\}$ такого, что не только средняя ошибка предикторов, построенных по отобранным признакам, минимальна, но и среднее попарное расстояние между этими предикторами в пространстве прогнозов максимально. Если эту задачу удастся решить, то подмножество $\{f\}$ будет содержать номера признаков, хорошо описывающих зависимости в данных, но при этом удаленных друг от друга, то есть разнородных, что должно повысить обобщающую способность окончательной модели и устойчивость к выбросам.

В процессе обучения нам доступны только данные из обучающей выборки, поэтому на практике будем решать задачу для тренировочных данных:

$$\frac{1}{|\mathbb{X}_{train}|}\sum_{(X,Y)\in(\mathbb{X}_{train},\mathbb{Y}_{train})}\mathcal{L}(G,X,Y,\{f\})\to\min_{\{f\}}. \tag{17}$$

Как было сказано выше, глобальный минимум в этой задаче может быть найден только методом полного перебора подмножеств $\{f\}$ множества номеров исходных признаков $F$, что неразумно на практике из-за больших временных затрат. Поэтому перейдем к поиску локального экстремума этой задачи жадным алгоритмом. Итак:

1. Удалим из рассмотрения все признаки, для которых ошибка предикторов, построенных по ним, превышает некоторый заданный порог.

2. Запустим итерационный процесс. На шаге 1 построим множество $\{f\}_1$, состоящее из номера $l$ признака, по которому построен предиктор, имеющий наименьшее расстояние в пространстве прогнозов до вектора значений целевой переменной на обучающей выборке:

$$l=\arg\min_j\frac{1}{|\mathbb{X}_{train}|}\sum_{(X,Y)\in(\mathbb{X}_{train},\mathbb{Y}_{train})}\rho(G_j(X_j),Y),$$
$$\{f\}_1=\{l\}.$$

3. Пусть на шаге $k$ построено множество номеров признаков $\{f\}_k$. На шаге $(k+1)$ выберем индекс

$$l=\frac{1}{|\mathbb{X}_{train}|}\arg\min_m\sum_{(X,Y)\in(\mathbb{X}_{train},\mathbb{Y}_{train})}\mathcal{L}(G,X,Y,\{f\}_k\cup m), \tag{18}$$

где

$$\mathcal{L}(G,X,Y,\{f\}_k\cup m)=\frac{1}{k+1}\sum_{i\in\{f\}_k\cup m}L(G_i(X_i),Y)-\frac{1}{2(k+1)^2}\sum_{i,j\in(\{f\}_k\cup m)^2}\rho(G_i(X_i),G_j(X_j)),$$

и добавим его ко множеству $\{f\}_k$:

$$\{f\}_{k+1}=\{f\}_k\cup l.$$

4. Если на шаге $K$ достигнуто ограничение на число итераций итераций, то считать окончательное множество номеров признаков $\{f\}$ равным $\{f\}_K$.

Заметим, что на практике часто лучший результат достигается не при минимальном значении функционала (15) на обучающей выборке, а при значительно большем числе отобранных признаков, которое задается искусственно.

## 4.2 Отбор с использованием кластеризации: вариант A

Этот раздел посвящен модификации алгоритма FAST [7], которая может применяться и для случая вещественнозначных признаков в признаковых описаниях объектов.

Будем использовать построенные выше предикторы $G_i(X_i)$. Приведем полный предлагаемый алгоритм в базовой версии, а затем предложим его модификацию. Итак:

1. Для всех имеющихся признаков вычислить расстояния в пространстве прогнозов от предикторов, построенных по ним, до вектора истинных значений целевой переменной: $\rho(G_i(\mathbb{X}_{train,i}), \mathbb{Y}_{train})$. Удалить из рассмотрения признаки, для которых вычисленное значение выше заданного порога.

2. Для всех оставшихся признаков вычислить попарные расстояния в пространстве прогнозов между соответствующими предикторами: $\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j}))$. Создать полный граф $H$, где каждой вершине соответствует один из оставшихся признаков. Каждому ребру между вершинами $(i, j)$ присвоить вес, равный вычисленному значению $\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j}))$, взятому с отрицательным знаком.

3. Построить минимальное остовное дерево $T$ графа $H$ по алгоритму Прима [6].

4. Получить множество деревьев (лес) $\{T_i\}_{i=1}^K$ с помощью удаления ребер из остовного дерева $T$ по следующему правилу: удалить ребро $(i, j)$, если выполнено условие:

$$|\rho(G_i(\mathbb{X}_{train,i}), \mathbb{Y}_{train}) - \rho(G_j(\mathbb{X}_{train,j}), \mathbb{Y}_{train})| > k \cdot \rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j})), \qquad (19)$$

   где $k$ — настраиваемый параметр алгоритма.

5. Составить итоговое подмножество признаков следующим образом. Из каждого дерева $T_i$ полученного леса выбрать один признак, которому соответствует наименьшее значение расстояния до вектора целевых переменных от предиктора, построенного по нему:

$$\{f\} = \{j_i\}_{i=1}^K,$$

$$j_i = \arg\min_{j \in T_i} \rho(G_j(\mathbb{X}_{train,j}), \mathbb{Y}_{train}), \forall i \in \{1, \ldots, K\}. \qquad (20)$$

Таким образом, этот алгоритм теперь может быть применен к данным с вещественнозначными признаками. Вместо симметрической неопределенности в качестве весов на ребрах графа в этой версии алгоритма используется расстояние в пространстве прогнозов между ответами предикторов, построенных по признакам, взятое с отрицательным знаком. Изменено также условие удаления ребра: ребро $(i, j)$ будет удалено, если расстояние в пространстве прогнозов между ответами предикторов, соответствующих признакам $i$ и $j$, меньше, чем модуль разности расстояний от этих предикторов до вектора истинных значений целевой переменной. В зависимости от параметра $k$ меняется количество удаляемых алгоритмом ребер, а,

следовательно, и количество отбираемых признаков. С такими значениями весов при построении остовного дерева будут оставлены ребра между теми признаками, расстояние между прогнозами предикторов для которых максимально, то есть эти признаки разнородны. Другое условие удаления ребра соответствует условию того, что ошибка выпуклой комбинации этих двух предикторов меньше, чем ошибка каждого из них.

## 4.3 Отбор с использованием кластеризации: вариант Б

Предложим ниже одну модификацию описанного выше алгоритма. Вместо шага 2 используем следующие действия:

2. Для всех оставшихся признаков вычислить попарные расстояния в пространстве прогнозов соответствующих предикторов: $\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j}))$. Создать полный граф $H$, где каждой вершине соответствует один из оставшихся признаков. Каждому ребру между вершинами $(i, j)$ присвоить вес, равный значению

$$\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j})) - |\rho(G_i(\mathbb{X}_{train,i}), \mathbb{Y}_{train}) - \rho(G_j(\mathbb{X}_{train,j}), \mathbb{Y}_{train})|, \qquad (21)$$

которое можно интерпретировать как «эффективность» выпуклой комбинации из предикторов, соответствующих признакам $i$ и $j$, взятую с отрицательным знаком.

Вместо шага 4 будем использовать следующее:

4. Получить множество деревьев (лес) $\{T_i\}_{i=1}^K$ с помощью удаления ребер из остовного дерева по следующему правилу: удалить ребро $(i, j)$, если выполнено условие:

$$\rho(G_i(\mathbb{X}_{train,i}), G_j(\mathbb{X}_{train,j})) - |\rho(G_i(\mathbb{X}_{train,i}), \mathbb{Y}_{train}) - \rho(G_j(\mathbb{X}_{train,j}), \mathbb{Y}_{train})| < p_\alpha, \qquad (22)$$

где $p_\alpha$ — $\alpha$-квантиль распределения весов остовного дерева, а $\alpha$ — настраиваемый параметр алгоритма.

Варьирование параметра $\alpha$ здесь позволяет изменять количество отбираемых признаков. Принципиальное отличие этой модификации алгоритма от базового — в том, что построение минимального остовного дерева происходит с максимизацией суммы «эффективностей» выпуклых комбинаций предикторов вместо максимизации расстояния между ними.

## 4.4 Комбинация методов

Возможно также использование предложенных методов в паре. Например, пусть после применения одной из модификаций алгоритма отбора признаков с помощью кластеризации получено множество индексов $\{f'\}$. Тогда применим алгоритм поиска оптимального признакового подпространства с использованием функционала ошибки выпуклой комбинации, считая множеством всех признаков множество $\{f'\}$. Подобные эксперименты будут описаны в следующих разделах.

# 5 Вычислительные эксперименты

Проведем несколько экспериментов, в которых сравним метод отбора признаков с помощью LASSO-регуляризации, алгоритм FAST, метод, использующий функционал ошибки выпуклой комбинации предикторов, а также две модификации метода, основанного на кластеризации признаков, и комбинированные методы. Протестируем все методы на нескольких наборах данных: данных о генотипе человека с целевой переменной, имеющей смысл роста человека, и данных об электрокардиограмме и фотоплетизмограмме человека с целевой переменной, имеющей смысл систолического кровяного давления.

Везде далее используем собственные реализации предложенных алгоритмов и алгоритма FAST на языке Python 3 с применением библиотек NumPy, SciPy, Scikit-learn, Matplotlib и некоторых других.

## 5.1 Задача предсказания роста человека

Рассмотрим задачу предсказания роста человека по данным о его генотипе. Имеющаяся выборка содержит данные о 784 людях, для каждого из которых известны значения 29682 признаков с возможным количеством значений не более 4-х. Вектор целевых переменных содержит информацию о росте людей. Значения целевой переменной колеблются в пределах от 150 до 200 сантиметров.

### 5.1.1 Регрессионная постановка

В регрессионной постановке этой задачи будем использовать в качестве критерия качества функцию отрицательного квадрата разности, которая при усреднении по всем объектам выборки приводит к среднеквадратичной ошибке MSE, взятой с отрицательным знаком:

$$S_{mean}(\mathbb{Y}_1, \mathbb{Y}_2) = -MSE(Y_1, Y_2) = -\frac{1}{N}\sum_{i=1}^{N}(\mathbb{Y}_{1,i} - \mathbb{Y}_{2,i})^2, \tag{23}$$

а также более информативный для нас обобщенный критерий — коэффициент детерминации $R^2$ [3]:

$$R^2(Y_{pred}, Y_{true}) = \frac{1 - SS_{res}}{1 - SS_{tot}}, \tag{24}$$

где $SS_{res}$ — сумма квадратов остатков регрессии:

$$SS_{res} = \sum_{i=1}^{N}(Y_{true,i} - Y_{pred,i})^2,$$

а $SS_{tot}$ — полная сумма квадратов:

$$SS_{tot} = \sum_{i=1}^{N}\left(Y_{true,i} - \frac{1}{N}\sum_{i=1}^{N}Y_{true,i}\right)^2.$$

Коэффициент детерминации показывает долю объясненной моделью дисперсии данных. Чем он ближе к оптимальному значению 1, тем лучше модель восстанавливает регрессию.

### 5.1.2 Классификационная постановка

Переформулируем задачу определения роста человека как задачу классификации с двоичными признаками. В первую очередь сделаем такое кодирование всех признаков, чтобы каждый из полученных признаков мог принимать только 2 различных значения. Затем найдем 0.5-квантиль распределения целевой переменной и всем объектам, имеющим значение целевой переменной, меньшее, чем эта квантиль, поставим в соответствие значение 0, а оставшимся — 1. Таким образом, получим новую выборку, содержащую 784 объекта и 60163 двоичных признака, а также бинарную целевую переменную со значениями 0 и 1. 0.5-квантиль распределения целевой переменной составляет 173.0 сантиметра.

В качестве обобщенного критерия качества будем использовать коэффициент AUC-ROC, определяемый как площадь под ROC-кривой [12]. Пусть есть модель, которая в качестве предсказания выдает вероятность принадлежности объектов к одному из двух классов. При необходимости привести ответ к двоичному виду встает вопрос о выборе порога бинаризации. Тогда ROC-кривая строится как ломаная через точки $(FPR_i, TPR_i)$ для всех возможных порогов бинаризации. Здесь $TPR_i$ (True Positive Rate) — доля объектов положительного класса, верно отнесенных классификатором к положительному классу, а $FPR_i$ (False Positive Rate) — доля объектов отрицательного класса, неверно отнесенных классификатором к положительному классу, при выборе $i$-го порога бинаризации.

## 5.2 Задача предсказания систолического давления

Рассмотрим еще одну задачу. Необходимо по данным электрокардиограммы (ЭКГ) и фотоплетизмограммы пациента определить его систолическое кровяное давление во время проведения этих измерений. У 832 пациентов эти исследования были проведены до определенного момента времени и еще у 495 — после. Данные о первых будем рассматривать как обучающую выборку, а данные о вторых — как валидационную. Затем к исходным признакам с использованием методики, основанной на построении оптимальных двумерных разбиений, были добавлены информативные признаки, и, таким образом, общая размерность признакового пространства составила 3491 признак, большая часть из которых принимает не более 4-х возможных значений. Вектор целевых переменных был преобразован в бинарный по следующему правилу: 0, если давление $<$ 125 мм рт. ст., и 1, если давление $>$ 140 мм рт. ст.

В качестве критерия качества будем использовать также коэффициент AUC-ROC, который был описан в предыдущем разделе.

## 5.3 Результаты экспериментов

Перейдем к подробному описанию параметров использовавшихся алгоритмов и полученным результатам.

В задаче определения роста человека имеющаяся выборка из 784 объектов была разбита на две равные части по 392 объекта для обучения и контроля случайным образом. Для оценки качества алгоритмов на обучении использовалась 10-фолдовая кросс-валидация на перемешанной обучающей выборке.

### 5.3.1 Задача предсказания роста человека в регрессионной постановке

Для решения задачи в регрессионной постановке в качестве базовых предикторов будем использовать стандартные модели Ridge-регрессии с коэффициентом регуляризации 1.0. В качестве фиксированных моделей, для которых будем оценивать качество работы на выборках с разной размерностью признакового пространства, выберем модели Ridge-регрессии с коэффициентом регуляризации 1.0 и эластичной сети (Elastic Net) [11] с коэффициентами общей регуляризации и доли $L_1$-регуляризации, равными 0.1.

Ниже приведем таблицу 1 результатов для задачи определения роста человека в регрессионной постановке. Столбец $|\{f\}|$ содержит размерность признакового пространства объектов, подаваемых на вход в регрессионные модели. Для контроля качества отбора признаков на обучении сначала составлялось подмножество отобранных признаков по всей обучающей выборке, а затем производилась кросс-валидация.

1. Сначала оценим среднеквадратичную ошибку MSE и коэффициент детерминации $R^2$ для указанных моделей на выборке, содержащей все 29682 признака.

2. Затем применим отбор признаков с помощью LASSO-регуляризации и оценим качество работы моделей на новой выборке, содержащей только признаки, имеющие коэффициент больше $10^{-5}$ в обученной модели LASSO-регрессии. Для этого будем использовать модель LASSO-регрессии, для которой подберем по кросс-валидации оптимальное значение коэффициента регуляризации, оказавшееся равным 0.457.

3. Далее применим отбор с использованием функционала. Для этого обучим 29682 модели Ridge-регрессии по каждому из признаков и отберем из них 5937 лучших по среднеквадратичной ошибке на обучающей выборке. Запустим алгоритм отбора по функционалу и оценим качество работы моделей на отобранных признаках.

4. На основе тех же 5937 лучших базовых предикторов применим отбор с помощью кластеризации в двух модификациях: модификацию с весами на ребрах графа, равными расстоянию в пространстве прогнозов между предикторами, будем обозначать как «кластеризация А», а другую — как «кластеризация Б». При этом будем выбирать наилучшие параметры $k$ и $\alpha$ алгоритмов соответственно по кросс-валидации на обучающей выборке. Оптимальное значение параметра $k$ и для кросс-валидации с фиксированной моделью Ridge-регрессии, и для кросс-валидации с фиксированной моделью Elastic Net оказалось равным 0.8. Оптимальное значение параметра $\alpha$ для Ridge-регрессии оказалось равным 0.05, а для Elastic Net — 0.3.

5. В заключение применим отбор по функционалу к подмножествам признаков, полученным после отбора с использованием кластеризации для параметров $k = 0.99$ и $\alpha = 0.25$ соответственно.

Поясним обозначения, используемые в таблицах:

- $|\{f\}|$ — *размерность признакового пространства после отбора признаков, произведенного на первом этапе;*
- $|\{f_{final}\}|$ — *размерность признакового пространства, непосредственно используемого фиксированной моделью при построении прогноза (используется отсечение по модулю коэффициента регрессии в линейных моделях и важности признака в случайном лесу с порогом* $10^{-8}$*);*
- *Ridge — стандартная реализация Ridge-регрессии из библиотеки Scikit-learn;*
- *LASSO — стандартная реализация LASSO-регрессии из библиотеки Scikit-learn;*
- *Elastic Net — стандартная реализация линейной регрессии с регуляризацией Elastic Net из библиотеки Scikit-learn;*
- *Logistic Regression — стандартная реализация логистической регрессии из библиотеки Scikit-learn;*
- *Random Forest — стандартная реализация решающего леса из библиотеки Scikit-learn;*
- *FAST — отбор признаков с помощью алгоритма FAST;*
- *Func, lim — отбор признаков с использованием функционала ошибки выпуклой комбинации и искусственным ограничением на число отбираемых признаков;*
- *Func, min — отбор признаков с использованием функционала ошибки выпуклой комбинации и выбором тех признаков, на которых достигается минимум функционала;*
- *Cluster A — отбор признаков с использованием кластеризации в базовой версии с расстояниями в пространстве прогнозов, взятыми с отрицательным знаком, на ребрах графа;*
- *Cluster B — отбор признаков с использованием кластеризации в модифицированной версии с «эффективностями» линейных комбинаций пар предикторов на ребрах графа.*
- *Cluster A/B + Func, lim/min — комбинированный метод, когда метод, использующий функционал ошибки выпуклой комбинации, применяется к признакам, отобранным одной из модификаций алгоритма, основанного на кластеризации.*

| Модель | Отбор | $\lvert\{f\}\rvert$ | $\lvert\{f_{final}\}\rvert$ | Train MSE | Train $R^2$ | Test MSE | Test $R^2$ |
|---|---|---|---|---|---|---|---|
| Ridge | | 29682 | 29390 | 72.999 | 0.231 | 79.353 | 0.320 |
| Elastic Net | | 29682 | 5413 | 70.903 | 0.254 | 74.747 | 0.359 |
| LASSO | | 29682 | 66 | 67.355 | 0.304 | 74.624 | 0.360 |
| Ridge | Func, lim | 400 | 400 | 57.360 | 0.445 | 65.278 | 0.440 |
| Elastic Net | Func, lim | 400 | 315 | 56.482 | 0.453 | **64.239** | **0.449** |
| Ridge | Func, min | 11 | 11 | 57.643 | 0.442 | 67.079 | 0.425 |
| Elastic Net | Func, min | 11 | 11 | 58.539 | 0.393 | 68.928 | 0.409 |
| Ridge | Cluster A | 5477 | 5477 | **23.657** | **0.771** | 100.804 | 0.136 |
| Elastic Net | Cluster A | 5477 | 2890 | 33.873 | 0.672 | 95.808 | 0.178 |
| Ridge | Cluster B | 298 | 298 | 57.904 | 0.440 | 68.601 | 0.412 |
| Elastic Net | Cluster B | 1782 | 701 | 48.986 | 0.526 | 73.875 | 0.367 |
| Ridge | Cluster B + Func, lim | 400 | 400 | 58.085 | 0.438 | 65.431 | 0.439 |
| Elastic Net | Cluster B + Func, lim | 400 | 318 | 56.884 | 0.449 | 64.756 | 0.445 |

Таблица 1: Задача предсказания роста человека. Регрессионная постановка. Отношение обучающей и контрольной выборок — 1:1. Оценка качества моделей

Проведем еще один эксперимент для оценки методов отбора признаков на тех же данных. Теперь будем делать кросс-валидацию с 5 случайными фолдами на всей выборке из 784 объектов следующим образом: на каждой итерации будем проводить отбор признаков по обучающей части в 80% выборки, а затем будем тестировать модели на тестовой части в 20% выборки с использованием только отобранных признаков. Результаты затем рассмотрим в виде средних и медианных значений.

| Модель | Отбор | $\lvert\{f\}\rvert$ | $\lvert\{f_{final}\}\rvert$ | Mean test $R^2$ | Median test $R^2$ |
|---|---|---|---|---|---|
| Ridge | | 29682 | 29450 | 0.343 | 0.336 |
| Elastic Net | | 29682 | 6503 | 0.356 | 0.346 |
| Ridge | Func, lim | 200 | 200 | 0.439 | 0.448 |
| Elastic Net | Func, lim | 200 | 178 | 0.443 | 0.443 |
| Ridge | Func, lim | 400 | 400 | 0.437 | 0.432 |
| Elastic Net | Func, lim | 400 | 297 | **0.455** | **0.462** |
| Ridge | Cluster B | 1485 | 1485 | 0.364 | 0.359 |
| Elastic Net | Cluster B | 1485 | 492 | 0.446 | 0.455 |

Таблица 2: Задача предсказания роста человека. Регрессионная постановка. Отношение обучающей и контрольной выборок — 4:1. Оценка качества моделей

### 5.3.2 Задача предсказания роста человека в классификационной постановке

В качестве базовых предикторов для решения этой задачи выберем модели логистической регрессии с коэффициентом регуляризации 0.1. В качестве фиксированных моделей для оценки качества будем использовать модель логистической регрессии с коэффициентом регуляризации 0.1, эластичную сеть для классификации с общим коэффициентом регуляризации 0.1 и долей $L_1$-регуляризации, равной 0.05, и случайный лес из 100 деревьев с минимальным количеством объектов в листе, равным 3.

Ниже приведем таблицу 3, содержащую оценки качества моделей, обученных на выборках, содержащих только отобранные признаки. Сравним разные методы отбора.

Как и в предыдущей постановке задачи, сначала оценим качество моделей на полной выборке, то есть в случае, когда отбор признаков не производится. Далее применим описанный выше алгоритм FAST, предложенные в этой работе его модификации, метод, использующий функционал ошибки выпуклой комбинации и оба предложенных метода в паре. Кросс-валидацию будем делать следующим образом: сначала по всей обучающей выборке определяется подмножество признаков, а затем модели оцениваются по результатам работы на 10 фолдах.

| Модель | Отбор | $\lvert\{f\}\rvert$ | $\lvert\{f_{final}\}\rvert$ | Train AUC-ROC | Test AUC-ROC |
|---|---|---|---|---|---|
| Logistic Regression | | 60163 | 59795 | 0.806 | 0.816 |
| Elastic Net | | 60163 | 4275 | 0.802 | 0.805 |
| Random Forest | | 60163 | 2833 | 0.821 | 0.828 |
| Logistic Regression | FAST | 195 | 195 | 0.838 | 0.817 |
| Elastic Net | FAST | 195 | 103 | 0.825 | 0.812 |
| Random Forest | FAST | 195 | 184 | 0.831 | **0.833** |
| Random Forest | Func, lim | 400 | 371 | 0.574 | 0.517 |
| Elastic Net | Func, min | 3 | 3 | 0.792 | 0.780 |
| Logistic Regression | Cluster A | 5711 | 5711 | **0.988** | 0.771 |
| Random Forest | Cluster A | 5672 | 2265 | 0.922 | 0.824 |
| Random Forest | Cluster B | 1806 | 1465 | 0.826 | 0.805 |
| Random Forest | Cluster A + Func, min | 3 | 3 | 0.769 | 0.777 |
| Random Forest | Cluster B + Func, min | 6 | 6 | 0.734 | 0.626 |
| Random Forest | Cluster A + Func, lim | 400 | 382 | 0.546 | 0.511 |

Таблица 3: Задача предсказания роста человека. Классификационная постановка. Оценка качества моделей

### 5.3.3 Задача предсказания систолического давления

Аналогичную вышеописанным процедуру применим в задаче предсказания систолического давления с той лишь разницей, что контрольная выборка в этой задаче полностью отделена от обучающей, поскольку собрана после определенного момента времени. В качестве базовых моделей используем так же логистическую регрессию, эластичную сеть для классификации и случайный лес.

Результаты приведем ниже в таблице 4.

| Модель | Отбор | $\lvert\{f\}\rvert$ | $\lvert\{f_{final}\}\rvert$ | Train AUC-ROC | Test AUC-ROC |
|---|---|---|---|---|---|
| Logistic Regression | | 3491 | 3491 | 0.941 | 0.782 |
| Elastic Net | | 3491 | 776 | 0.932 | 0.777 |
| Random Forest | | 3491 | 2159 | 0.946 | **0.806** |
| Random Forest | Func, min | 6 | 6 | 0.912 | 0.685 |
| Random Forest | Func, lim | 400 | 398 | **0.946** | 0.766 |
| Random Forest | Cluster A | 426 | 419 | 0.941 | 0.793 |
| Random Forest | Cluster B | 874 | 814 | 0.945 | 0.787 |
| Logistic Regression | Cluster A + Func, lim | 400 | 400 | 0.934 | 0.722 |
| Elastic Net | Cluster A + Func, lim | 400 | 197 | 0.930 | 0.733 |
| Random Forest | Cluster A + Func, lim | 400 | 396 | 0.940 | 0.784 |
| Logistic Regression | Cluster B + Func, lim | 400 | 400 | 0.931 | 0.726 |
| Elastic Net | Cluster B + Func, lim | 400 | 198 | 0.933 | 0.728 |
| Random Forest | Cluster B + Func, lim | 400 | 399 | 0.939 | 0.789 |

Таблица 4: Задача предсказания систолического давления. Оценка качества моделей

## 5.4 Обсуждение и выводы

Таким образом, были проведены эксперименты с существующими и предложенными методами оптимизации признакового пространства на имеющихся данных в задачах предсказания роста человека по данным о его генотипе и предсказания систолического давления по данным электрокардиограммы и фотоплетизмограммы.

Эксперименты показали, что в регрессионной постановке задачи предсказания роста предложенные методы отбора признаков с использованием функционала ошибки выпуклой комбинации и с использованием кластеризации, а также их комбинация позволяют повысить коэффициент детерминации $R^2$ со значения 0.36, достигаемого эластичной сетью и LASSO-регрессией, до значений 0.45 – 0.46, то есть примерно на 25%. При этом удается сократить размерность признакового пространства в несколько десятков раз.

В классификационной постановке задачи предсказания роста человека и в задаче предсказания систолического давления с помощью кластеризационных, а также комбинированных методов удается достичь сокращения размерности признакового пространства примерно в

10 – 12 раз при снижении коэффициента AUC-ROC не более, чем на 1 – 2%. Однако для метода, использующего функционал, в этих задачах не удается получить подобного результата.

Заметим, что предложенные методы не основываются на предположении о дискретной природе признаков или целевой переменной, поэтому позволяют решать задачу оптимизации признакового пространства в более широкой постановке.

# 6 Заключение

Итак, в данной работе были предложены два метода решения задачи оптимизации признакового пространства. Оба метода используют базовые предикторы (такие, как Ridge-регрессоры или модели логистической регрессии), обученные только по данным об одном признаке объектов, и преследуют цель такого выбора подпространства признаков, что каждый из признаков в нем хорошо коррелирует с целевой переменной, но при этом они сильно удалены друг от друга в пространстве прогнозов базовых предикторов, построенных по ним. В работе были проведены исследования качества работы предложенных методов в регрессионных и классификационных задачах предсказания роста человека по данным о его генотипе и предсказания систолического давления по данным ЭКГ и фотоплетизмограммы человека. Поставленные эксперименты показали, что с помощью предложенных методов удается добиться значительного улучшения качества модели, построенной по полученному подпространству признаков, в регрессионной задаче и значительного сокращения размерности признакового пространства при незначительном снижении качества моделей в классификационных задачах.

# Список литературы


[1] Н. Ю. Хомутов. Методы повышения эффективности моделей машинного обучения, основанные на различных принципах снижения размерности. Магистерская диссертация. МГУ, ВМК, каф. ММП. 2017.

[2] А.А. Докукин, О.В. Сенько, Н.Н. Киселева, and Н. Ю. Хомутов. Двухуровневый метод построения линейных регрессий с использованием наборов оптимальных выпуклых комбинаций. *Доклады Академии наук*, 479:11–13, 2018.

[3] Richard Anderson-Sprecher. Model comparisons and r2. *American Statistician*, 48(2):113–117, 1994. ISSN 15372731. doi: 10.1080/00031305.1994.10476036.

[4] Leo Breiman. Random Forests. pages 1–33, 2001. ISSN 1098-6596. doi: 10.1017/CBO9781107415324.004.

[5] Heewon Park, Seiya Imoto, and Satoru Miyano. Recursive random lasso (RRLasso) for identifying anti-cancer drug targets. *PLoS ONE*, 10(11):1–19, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0141869.

[6] R. C. Prim. Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, 36(6):1389–1401, 1957. ISSN 15387305. doi: 10.1002/j.1538-7305.1957.tb01515.x.

[7] Qinbao Song, Jingjie Ni, and Guangtao Wang. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):1–14, 2013. ISSN 1041-4347. doi: 10.1109/TKDE.2011.181. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5989810`.

[8] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso Robert Tibshirani. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1):267–288, 1996. ISSN 0035-9246. doi: 10.1111/j.1467-9868.2011.00771.x.

[9] N S Tzannes and J P Noonan. The mutual information principle and applications. *Information and Control*, 22(1):1–12, 1973. URL `http://www.scopus.com/inward/record.url?eid=2-s2.0-0015588092&partnerID=40`.

[10] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24(1):175–186, 2014. ISSN 09410643. doi: 10.1007/s00521-013-1368-0.

[11] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2):301–320, 2005. ISSN 13697412. doi: 10.1111/j.1467-9868.2005.00503.x.

[12] M. H. Zweig and G. Campbell. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577, 1993. ISSN 00099147. doi: ROC; Receiver-Operating Characteristic; SDT; Signal Detection Theory.