Федеральное государственное автономное образовательное учреждение
высшего образования «Московский физико-технический институт
(государственный университет)»
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

На правах рукописи
УДК 519.7

Зухба Анастасия Викторовна

ОЦЕНКА ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ ЗАДАЧ ОТБОРА ЭТАЛОННЫХ ОБЪЕКТОВ И ПРИЗНАКОВ

Специальность 01.01.09 —
«Дискретная математика и математическая кибернетика»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук
профессор РАН
Воронцов Константин Вячеславович

Долгопрудный — 2018

Оглавление

Стр.

Введение

Диссертационная работа посвящена проблеме отбора объектов и признаков для построения метрических и монотонных классификаторов. В работе предложены оптимизационные постановки задач отбора объектов и признаков, проведена оценка их вычислительной сложности. Предложен приближенный алгоритм монотонизации с одновременным отбором объектов и признаков. Проведена эксперементальная проверка алгоритма на данных задачи медицинской диагностики.

Актуальность темы исследования и степень её разработанности

Широкое распространение компьютерных технологий для сбора и накопления данных практически во всех областях жизни стимулирует развитие методов интеллектуального анализа данных, предсказательного моделирования, машинного обучения [1; 2]. В частности, для автоматизации принятия решений широко используются модели, методы и алгоритмы распознавания образов или классификации. Методы машинного обучения часто представляют собой приближенные решения NP-трудных задач [3]. На практике к этим методам предъявляются требования высокой обобщающей (предсказательной) способности и низкой вычислительной сложности.

Задача классификации заключается в том, чтобы по заданному конечному множеству объектов, разделенных на классы, построить функцию классификации, которая произвольному объекту той же природы ставит в соответствие один из заданных классов. Например, в задачах медицинской диагностики объектами являются признаковые описания состояния человека; в роли признаков могут выступать симптомы, результаты обследований или биомаркеры; классы соответствуют диагностируемым заболеваниям.

Большинство моделей классификации явно или неявно формализуют гипотезу компактности — неформальное предположение о том, что схожие объекты чаще принадлежат одному классу, чем разным. Успех в решении задачи классификации во многом зависит от того, насколько адекватно применяемая модель классификации выражает понятие «сходства» объектов.

При решении практических задач классификации часто возникает необходимость отбора объектов и признаков. Отбор признаков (feature selection,

FS) необходим для выявления информативных подпространств признаков, в которых выполняется гипотеза компактности. Отбор объектов (prototype selection, PS) необходим для отсева ошибочных объектов (выбросов) и выявления типичных представителей классов (эталонов), достаточных для понимания структуры класса и надёжной классификации остальных объектов. Кроме того, отбор как объектов, так и признаков, позволяет сокращать объём хранимых данных, уменьшать время обучения алгоритма, повышать обобщающую способность и устойчивость классификации. Задачи отбора объектов и признаков в литературе, как правило, рассматриваются по отдельности. Редкое исключение составляют работы новосибирской школы распознавания образов (Н. Г. Загоруйко, Г. С. Лбов, И. А. Борисова, В. В. Дюбанов, О. А. Кутненко и др.). Целесообразность единого алгоритмического решения этих двух задач следует из того факта, что результат отбора объектов может зависеть от признакового пространства, а результат отбора признаков может зависеть от способа отсева выбросов или выделения эталонов.

Возможность отбора эталонов наиболее естественно возникает в метрических методах классификации, поскольку для них понятие «сходства» объектов формализуется в явном виде.

Для построения метрических алгоритмов классификации существуют различные эвристические методы отбора эталонных объектов [4–6], в том числе основанные на минимизации частоты ошибок на обучающей выборке.

В [7; 8] М. Н. Ивановым и К. В. Воронцовым предложены алгоритмы, основанные на минимизации функционала полного скользящего контроля. В этом методе явная оптимизация обобщающей способности позволяет улучшать множество отбираемых эталонов.

Во всех перечисленных методах применяются жадные стратегии, не гарантирующие оптимальности решения, то есть что множество эталонов будет иметь минимальную мощность и/или обеспечивать минимальное значение критерия качества. Многие подходы не предполагают постановку задачи как оптимизационной. Вместо этого предлагается алгоритм отбора эталонов, качество которого исследуется «пост фактум» чисто эмпирически. Оптимизационные постановки задачи отбора эталонов и признаков, а также вопросы их вычислительной сложности, являются относительно малоисследованными.

В данной работе рассматриваются оптимизационные постановки задач отбора объектов и признаков в метрических классификаторах. Кроме того, данная задача естественным образом обобщается на случай монотонных классификаторов. Предположение о монотонности функции классификации возникает во многих прикладных задачах, и его явный учёт позволяет повышать обобщающую способность метода классификации [9–12]. В данной работе рассматривается полный набор оптимизационных постановок задач отбора объектов и признаков для построения монотонных классификаторов.

На практике часто пользуются линейными моделями классификации с неотрицательными коэффициентами, как самым простым способом реализации монотонного классификатора. Преимущество линейной модели — в её простоте и наличии готовых реализаций. Однако для многих задач линейная модель представляется слишком жёсткой. Множество монотонных функций существенно шире, чем множество линейных функций с неотрицательными коэффициентами. Поэтому нелинейные монотонные модели классификации имеют преимущества в задачах со сложной разделяющей поверхностью, что было показано на примерах задач медицинской диагностики [13; 14], ранжирования поисковой выдачи [15], категоризации текстов [16], при построении композиций классификаторов [17; 18].

Известно много методов построения монотонных классификаторов [12;14; 19–21]. Некоторые функции расстояний [22–24] для метрических алгоритмов классификации позволяют строить метрические монотонные классификаторы.

Для монотонного метода ближайшего соседа [22] имеются высокоточные комбинаторные оценки полного скользящего контроля [23;24], из которых следует, что данный метод обладает высокой обобщающей способностью благодаря максимизации ширины зазора между классами. Данный факт позволяет провести аналогии между монотонными классификаторами и методом опорных векторов SVM [25], который является одним из лучших методов классификации именно благодаря принципу максимизации зазора.

Большинство методов монотонной классификации требуют предварительной монотонизации выборки, которая выполняется с помощью жадных эвристических алгоритмов [26] или методов изотонной регрессии [27]. Монотонизация сводится к отбрасыванию объектов обучающей выборки, нарушающих условие монотонности, и поиску пространства признаков, в котором условия

монотонности выполняются. Некоторые подходы к решению задачи монотонизации основаны на изменении известных классов объектов [28; 29]. Сложность построения монотонного классификатора оценивается снизу вычислительной сложностью задачи монотонизации выборки.

Как и в случае с метрическими алгоритмами, вопросы оптимизационной постановки задачи монотонизации и их вычислительной сложности являются малоисследованными.

Цели и задачи

Целью данной работы является выяснение статуса вычислительной сложности задач отбора эталонных объектов и признаков для метрических и монотонных классификаторов. Для достижения поставленной цели решаются следующие задачи:

1. Оценить вычислительную сложность оптимизационных постановок задач отбора объектов и признаков для алгоритма ближайшего соседа.
2. Оценить вычислительную сложность оптимизационных постановок задачи монотонизации обучающей выборки.
3. Разработать алгоритм монотонизации с одновременным отбором объектов и признаков.

Научная новизна

1. Получена оценка вычислительной сложности задачи отбора объектов и признаков для алгоритма ближайшего соседа.
2. Предложена систематизация оптимизационных постановок задачи монотонизации выборки.
3. Получена оценка вычислительной сложности задачи монотонизации выборки.
4. Предложен и протестирован экспериментально алгоритм монотонизации выборки с одновременным отбором объектов и признаков.

Теоретическая и практическая значимость работы

Оптимизационная постановка позволяет получить оценки вычислительной сложности задач отбора объектов и признаков при различных целевых функциях и ограничениях. Систематизация получаемых задач дискретной оптимизации позволяет выбирать целевые функции, которые соответствуют ре-

шаемой прикладной задаче. Доказательство NP-полноты обосновывает применение субоптимальных эвристических методов для решения соответствующих оптимизационных задач. Предложенный в работе приближенный алгоритм монотонизации с одновременным отбором объектов и признаков частично решает проблему «застревания» в локальных минимумах, связанных с шумовыми объектами и неинформативными признаками. Для единственной постановки задачи, имеющей полиномиальную сложность, указан точный эффективный алгоритм решения.

Методология и методы исследования

Для анализа постановок задач отбора объектов и признаков использовались элементы комбинаторики и теории графов. При оценке вычислительной сложности использовались методы сведения классических задач дискретной оптимизации (задачи о биклике, задачи о покрытии множеств подмножествами, задачи о вершинном покрытии) к задачам отбора объектов и признаков при обучении классификации. В целях проверки предложенного алгоритма монотонизации был проведен вычислительный эксперимент на прикладной задаче информационного анализа электрокардиосигналов.

Положения, выносимые на защиту

1. Оценка вычислительной сложности задачи отбора признаков и эталонных объектов для алгоритма ближайшего соседа.
2. Оценки вычислительной сложности задачи монотонизации обучающей выборки в различных постановках.
3. Приближенный алгоритм монотонизации обучающей выборки с одновременным отбором объектов и признаков.

Степень достоверности и апробация результатов

Достоверность теоретических результатов обеспечивается математическими доказательствами теорем. Результаты экспериментов соответствуют результатам, полученным другими авторами [30].

Основные результаты работы докладывались на следующих научных конференциях:

– 52-я научная конференция МФТИ, Москва–Долгопрудный 2009. [31]
– 53-я научная конференция МФТИ, Москва–Долгопрудный 2010. [32]

– Всероссийская конференция «Математические методы распознавания образов», ММРО-15, Петрозаводск 2011. [33]
– Всероссийская конференция «Математические методы распознавания образов», ММРО-16, Казань 2013. [34]
– Всероссийская конференция «Математические методы распознавания образов», ММРО-17, Светлогорск 2015. [35; 36]
– Всероссийская конференция «Математические методы распознавания образов», ММРО-18, Таганрог 2017. [37]

Публикации

[38–40]

Основные результаты по теме диссертации изложены в десяти публикациях [31–40], две из которых опубликованы в изданиях из перечня, рекомендованного ВАК [38; 39], семь — в тезисах докладов [31–37].

Личный вклад автора а публикации с соавторами заключался в разработке и обосновании различных версий алгоритмов монотонизации и подготовке текста публикации.

Глава 1. Обучение по прецедентам как задача оптимизации

Пусть имеется множество объектов $\mathbb{X}$ и множество ответов Y, и существует функция $y\colon \mathbb{X} \to Y$, значение для которой известно только на конечном подмножестве $\{x_1, \ldots, x_L\} \subset \mathbb{X}$.

Опр. 1.0.1. *Известные пары «объект–ответ»* (x_i, y_i) *называются прецедентами, а совокупность* $X^L = (x_i, y_i)_{i=1}^L$, *где* $y_i = y(x_i)$, *— обучающей выборкой.*

Задача обучения по прецедентам состоит в том, чтобы восстановить функциональную зависимость между объектами и ответами, то есть построить отображение $\gamma\colon \mathbb{X} \to Y$, которое аппроксимирует функцию $y(x)$.

Опр. 1.0.2. *Отображения* $\gamma\colon \mathbb{X} \to Y$, *аппроксимирующее функцию* $y(x)$ *будем называть алгоритмами.*

Опр. 1.0.3. *Моделью алгоритмов называется параметрическое семейство отображений* $\Gamma_\Theta = \{\gamma_\Theta(x, \theta) | \theta \in \Theta\}$, *где* $\gamma_\Theta\colon \mathbb{X} \times \Theta \to Y$, *а* Θ *— множество допустимых параметров* θ.

Опр. 1.0.4. *Методом обучения (learning algorithm) называется отображение* μ, *которое произвольной конечной выборке* X^L *ставит в соответствие алгоритм* $\gamma\colon \mathbb{X} \to Y$.

Как правило, алгоритм выбирается из некоторого параметрического семейства Γ_Θ, а метод обучения μ является методом решения оптимизационной задачи выбора параметра θ из множества Θ. Целевые функции рассматриваемых оптимизационных задач зависят от параметрического семейства Γ_Θ и метода обучения μ. Примеры таких целевых функций, каждую из которых можно рассматривать как оценку качества классификатора, приведены в разделе 1.1.

1.1 Измерение качества классификаторов и обобщающая способность

Опр. 1.1.1. *Индикатором ошибки алгоритма γ на объекте x_i называется функция, принимающая значение 0, если ответ алгоритма совпадает с истинным ответом, и 1 в противном случае:*

$$I(x_i, \gamma(x_i)) = \big[y(x_i) \neq \gamma(x_i)\big].$$

Опр. 1.1.2. *Частота ошибок алгоритма γ на выборке X^L определяется как*

$$\nu(\gamma, X^L) = \frac{1}{L}\sum_{i=1}^{L} I(x_i, \gamma(x_i)).$$

Опр. 1.1.3. *Если $\nu(\gamma, X) = 0$, говорят что алгоритм γ корректно работает на множестве X.*

Частота ошибок является одной из самых простых целевых функций, используемых при обучении алгоритмов. Ее недостатком является то, что малая частота ошибок на обучающей выборке еще не гарантирует, что построенный алгоритм будет также редко ошибаться на новых (контрольных) объектах.

Алгоритм обучения обладает *обобщающей* способностью (generalization ability), если вероятность ошибки на контрольных объектах достаточно мала или хотя бы предсказуема, то есть не сильно отличается от ошибки на обучающей выборке. Если вероятность ошибки обученного алгоритма на контрольных объектах оказывается существенно выше, чем средняя ошибка на обучающей выборке, говорят что произошло *переобучение* [41–43]. Переобучение возникает при использовании избыточно сложных моделей.

Для оценивания обобщающей способности метода используются функционалы качества, основанные на принципе скользящего контроля.

Пусть дана выборка X^L длинны L. Разобьём её на два непересекающихся подмножества: обучающую подвыборку (training set) X^ℓ и контрольную подвыборку (testing set) X^k, где $L = k + \ell$.

Обозначим через $(X_n^\ell, X_n^k),\ n = 1, \ldots, N$ всевозможные разбиения выборки X^L на обучающую и контрольную подвыборки, $N = C_L^\ell$.

Опр. 1.1.4. *Функционалом полного скользящего контроля (complete cross validation, CCV) называется средняя частота ошибок на контрольных подвыборках [44]:*

$$Q_k(\mu) = \frac{1}{N}\sum_{n=1}^{N} \nu(\mu(X_n^\ell), X_n^k).$$

При $k = 1$ функционал CCV переходит в другой известный функционал — скользящий контроль с одним отделяемым объектом (leave-one-out, LOO):

$$Q_1(\mu) = \frac{1}{L}\sum_{i=1}^{L} \nu(\mu(X^L \backslash \{x_i\}), \{x_i\}).$$

1.2 Задачи отбора объектов и признаков

Опр. 1.2.1. *Признаком f называют отображение $f : \mathbb{X} \to D_f$, где D_f — множество допустимых значений признака.*

Опр. 1.2.2. *Пусть дан набор признаков $\mathbb{F} = \{f_1, \dots f_n\}$. Вектор $\big(f_1(x), \dots f_2(x)\big)$ называют* признаковым описанием *объекта x.*

Как правило объекты задаются своими признаковыми описаниями.

Необходимым условием существования решения задачи классификации является *гипотеза компактности* [45]. Гипотеза компактности заключается в том, что если мера сходства объектов введена достаточно удачно, схожим объектам очень часто соответствуют схожие ответы.

Представление о сходстве между объектами строится по их признаковому описанию. То есть выполнение гипотезы компактности зависит от рассматриваемого пространства признаков. Признаки, от которых не зависят классы объектов, могут мешать выполнению гипотезы компактности или способствовать переобучению. Таким образом возникает задача отбора признаков.

Теперь предположим, что мера сходства введена достаточно удачно. Чаще всего реальные данные являются *зашумленными* — то есть среди объектов встречаются нетипичные представители классов, лежащие ближе к объектам чужого класса, чем своего, то есть нарушающие гипотезу компактности. Такие

объекты могут возникать как следствие ошибки измерений признаков, ошибки классификации или быть следствием вероятностной природы задачи. Изъятие таких нетипичных объектов из обучающей выборки увеличивает обобщающую способность классификаторов. Как следствие, возникает задача отбора объектов. Существуют и другие случаи, когда отбор объектов полезен: например, для сокращения объема хранимой информации или для выявления структуры класса.

Задача отбора признаков, как и задача отбора объектов, представляет собой задачи дискретной оптимизации. Для удобства обозначим FS — отбор признаков, PS — отбор объектов.

1.3 Вычислительная сложность некоторых задач дискретной оптимизации

Для выяснения статуса вычислительной сложности рассматриваемых в данной работе задач использовались следующие известные задачи дискретной оптимизации [46–48].

Задача о покрытии множества подмножествами. *Покрытием* конечного множества U семейством его подмножеств S называется такое множество подмножеств $C = \{C_1, \ldots, C_k\} \subseteq S$, что $\bigcup\limits_{i=1}^{k} C_i = U$. *Размером* покрытия C называют число $k = |C|$.

Задача 1.3.0.1. *Найти покрытие $C \subseteq S$ множества U минимального размера k.*

Данная задача является NP-полной.

Обычно предполагают, что покрытие существует, то есть $\bigcup\limits_{C_i \in S} C_i = U$. Задача имеет еще одну формулировку, также являющуюся NP-полной: существует ли покрытие $C \subseteq S$ множества U размера $k \leq k_0$, где k_0 — заданное число.

Задача о биклике. *Полным двудольным графом* или *бикликой* называется двудольный граф, у которого любая вершина первой доли соединена со всеми вершинами второй доли вершин.

Задача 1.3.0.2. *Дан двудольный граф. Найти полный двудольный подграф $K_{i,i}$ с максимальным числом рёбер i^2.*

Данная задача является NP-полной.

Задача о о вершинном покрытии графа. *Вершинное покрытие* неориентированного графа $G = (V, E)$ — это множество его вершин S, такое, что у каждого ребра графа хотя бы один из концов принадлежит S.

Задача 1.3.0.3. *Дан граф $G = (V, E)$. Найти его вершинное покрытие S минимального размера $|S|$.*

Данная задача является NP-полной.

Задача о вершинном покрытии в двудольном графе. *Паросочетанием* называют множество попарно несмежных ребер, то есть рёбер, не имеющих общих вершин. *Максимальное паросочетание* — это такое паросочетание в графе, которое не содержится ни в каком другом паросочетании этого графа.

Задача 1.3.0.4. *Дан двудольный граф $G = (V, E)$. Найти его вершинное покрытие S минимального размера $|S|$.*

Точное решение данной задачи вычисляется за полиномиальное время.

Теорема 1.3.1. *(Кёниг) Мощность максимального паросочетания в двудольном графе равна мощности его минимального вершинного покрытия.*

Доказательство этой теоремы конструктивно: строится алгоритм, находящий вершинное покрытие двудольного графа G за время $O\big(|E|\sqrt{|L_G| + |R_G|}\big)$, где L_G, R_G — правая и левая доли соответственно [49].

1.4 Основные выводы главы 1

Построение алгоритма классификации представляет собой решение оптимизационной задачи поиска приближения функции $y\colon \mathbb{X} \to Y$. Её решение можно условно разбить на несколько этапов:

- отбор признаков,
- построение функции расстояний,
- отбор объектов,
- оптимизация параметров алгоритма из выбранного параметрического семейства.

Данные этапы решения могут производится в различном порядке, вместе или отдельно. Каждый из них является решением соответствующий оптимизационной задачи.

Глава 2. Отбор эталонов и признаков для метрических классификаторов без ограничений монотонности

Метрическими называются алгоритмы классификации, основанные на измерении сходства между парами объектов. Типичным представителем метрических алгоритмов является алгоритм ближайшего соседа.

Метод обучения μ для классификации по ближайшему соседу (Nearest Neighbor classifier, NN) сводится к тривиальному запоминанию обучающей выборки. После этого произвольный классифицируемый объект $u \in \mathbb{X}$ относится к тому классу, которому принадлежит ближайший к нему обучающий объект. Для формализации понятия близости (сходства) на $\mathbb{X}$ вводится функция расстояния $\rho(x,x')$, вообще говоря, не обязательно метрика. Как уже говорилось, выполнение гипотезы компактности является обязательным условием для существования решения задачи классификации.

В данной главе приводится формализация гипотезы компактности. С точки зрения приведенной формализации рассматриваются оптимизационные постановки задачи отбора объектов и признаков. Оценивается вычислительная сложность полученных постановок задач отбора объектов и признаков.

2.1 Связь обобщающей способности классификатора ближайшего соседа и гипотезы компактности

Для произвольного $x_i \in X^L$ положим $x_i \equiv x_{i0}$ и обозначим через $x_{i1}, \ldots, x_{i,L-1}$ последовательность всех объектов выборки X^L, упорядоченную по возрастанию расстояний $\rho(x_i, x_{ij}),\ j = 0, \ldots, L-1$.

Обозначим через $r_m(x_i)$ ошибку, возникающую при замене известной классификации объекта x_i на ответ $y(x_{im})$ на m-ом соседе, то есть $r_m(x_i) = I(x_i, y(x_{im}))$.

Опр. 2.1.1. *Профилем компактности [7] выборки X^L называется функция $P(m)$, выражающая долю объектов выборки, для которых правильный ответ*

не совпадает с правильным ответом на m-ом соседе:

$$P(m) = \frac{1}{L}\sum_{i=1}^{L} r_m(x_i); \quad m = 1,\ldots,L-1.$$

Будем говорить, что профиль компактности $P(m)$ оценивает компактность выборки. Высокие значения профиля компактности $P(m)$ для первых нескольких значений m означают, что гипотеза компактности на выборке X^L с заданной функцией расстояний ρ выполняется.

Следующая теорема позволяет эффективно вычислять функционал полного скользящего контроля для метода ближайшего соседа.

Теорема 2.1.1 (Воронцов [11]). *Для задачи классификации методом ближайшего соседа справедливо следующее точное выражение функционала полного скользящего контроля Q_k:*

$$Q_k(\mu) = \sum_{m=1}^{k} P(m)C(m),$$

где $C(m) = C_{L-1-m}^{\ell-1}/C_{L-1}^{\ell}$.

Коэффициенты $C(m)$ не зависят от самой выборки X^L и являются постоянными при известных L, k, $\ell = L - k$, m. Таким образом функционал качества Q_k, оценивающий обобщающую способность алгоритма ближайшего соседа, можно рассматривать еще и как оценку компактности выборки X^L.

Теперь рассмотрим более сложный метод обучения μ_Ω, который запоминает не всю обучающую выборку, а лишь подмножество эталонных объектов $\Omega \subseteq X^L$. На стадии классификации используется тот же алгоритм ближайшего соседа, но теперь ближайшие соседи выбираются только из Ω. Обозначим этот алгоритм через γ_Ω.

Естественным критерием для отбора эталонов Ω является минимум частоты ошибок на всех остальных объектах:

$$\nu(\gamma_\Omega, X^L\backslash\Omega) \to \min_{\Omega}. \tag{2.1}$$

Однако такой способ отбора эталонов, как и всякая оптимизация параметров алгоритма по конечной выборке, может приводить к переобучению, вследствие

которого качество алгоритма γ_Ω будет хуже вне выборки X^L, которая в данном случае играла роль обучающей.

Рассмотрим возможные модификации функционала скользящего контроля, характеризующего обобщающую способность отбираемого множества эталонов Ω.

Вариант 1. Будем разбивать X^L на X_n^ℓ и X_n^k так, чтобы множество эталонных объектов Ω всегда находилось в X_n^ℓ, и ближайший сосед выбирался только из Ω. Число таких разбиений $N = C^k_{L-|\Omega|}$. Нетрудно убедиться, что данная модификация функционала скользящего контроля эквивалентна (2.1):

$$Q_k^*(\mu_\Omega) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{k}\sum_{x\in X_n^k} I(x,\gamma_\Omega(x)) = \frac{1}{Nk}\underbrace{\sum_{x\in X^L\setminus\Omega} I(x,\gamma_\Omega(x))}_{(L-|\Omega|)\nu(\gamma_\Omega,X^L\setminus\Omega)}\underbrace{\sum_{n=1}^{N}[x\in X_n^k]}_{C^{k-1}_{L-1-|\Omega|}} =$$

$$= \frac{C^{k-1}_{L-1-|\Omega|}}{C^k_{L-|\Omega|}}\frac{L-|\Omega|}{k}\,\nu(\gamma_\Omega,X^L\setminus\Omega) = \nu(\gamma_\Omega,X^L\setminus\Omega).$$

Вариант 2. Теперь разрешим эталонным объектам из Ω попадать как в обучение, так и в контроль. Потребуем, чтобы длина контрольной выборки была меньше числа эталонных объектов, чтобы гарантировать $X_n^\ell \cap \Omega \neq \varnothing$.

Обозначим через $r_m^\Omega(x_i)$ ошибку, возникающую при замене известной классификации объекта x_i на ответ $y(x_{im})$ на m-м соседе, $r_m^\Omega(x_i) = I(x_i,y(x_{im}))$, где x_{im} — m-ый объект из множества эталонов Ω, если упорядочить их по возрастанию расстояний до объекта x_i. Обратим внимание, что если $x_i \in \Omega$, то $m = 1,\ldots,|\Omega|-1$, а если $x_i \in X^L\setminus\Omega$, то $m = 1,\ldots,|\Omega|$.

Профилем Ω-компактности выборки X^L называется функция $P^\Omega(m)$, выражающая долю объектов выборки, для которых правильный ответ не совпадает с правильным ответом на m-ом соседе из множества эталонов:

$$P^\Omega(m) = \frac{1}{L}\sum_{i=1}^{L} r_m^\Omega(x_i).$$

Теорема 2.1.2. *Для задачи классификации методом ближайшего соседа справедливо следующее точное выражение функционала* $Q_k(\mu_\Omega)$*:*

$$Q_k(\mu_\Omega) = \sum_{m=1}^{k} P^\Omega(m) C(m).$$

Доказательство. Запишем в функционале скользящего контроля частоту ошибок через сумму индикаторов ошибки и переставим знаки суммирования:

$$Q_k(\mu_\Omega) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{k}\sum_{x\in X_n^k} I(x,\mu(X_n^\ell)) = \frac{1}{k}\sum_{i=1}^{L}\frac{1}{N}\underbrace{\sum_{n=1}^{N}[x_i \in X_n^k] I(x_i,\mu(X_n^\ell))}_{N_i}.$$

Внутренняя сумма, обозначенная фигурной скобкой, выражает число разбиений выборки X^L, при которых объект x_i оказывается в контрольной выборке и алгоритм $\mu(X_n^\ell)$ допускает на нем ошибку. Данная ситуация реализуется для таких разбиений, при которых m первых объектов из последовательности $x_{i0},x_{i1},\ldots$ попадают в контрольную подвыборку, а m-ый сосед находится в обучающей подвыборке и принадлежит другому классу. Причем в качестве соседей рассматриваются только объекты из множества Ω. Число таких разбиений в точности равно

$$N_i = \sum_{m=1}^{k} r_m^\Omega(x_i) C_{L-1-m}^{\ell-1},$$

поскольку $C_{L-1-m}^{\ell-1}$ есть число способов выбрать $(\ell-1)$ обучающих объектов из $X^L \setminus \{x_{i0},x_{i1},\ldots,x_{im}\}$. Тогда, используя определение Ω-профиля компактности и вынося общие множители, получим требуемое выражение.

Функционалы $Q_k(\mu_\Omega)$ и $Q_k^*(\mu_\Omega)$ тоже можно рассматривать как оценку компактности выборки X^L. В отличии от Q_k значения $Q_k(\mu_\Omega)$ и $Q_k^*(\mu_\Omega)$ зависят не только от компактности выборки, но и от того, насколько точно множество Ω описывает структуру классов.

2.2 Вычислительная сложность задачи отбора эталонов

Рассмотрим задачи минимизации по Ω функционала полного скользящего контроля $Q_k(\mu_\Omega)$ и его модификации $Q_k^*(\mu_\Omega)$ при заданной функции расстояния ρ. Далее будет показано, что обе задачи являются NP-трудными.

Доказательство NP-трудности будет выполнено путём сведения известной NP-полной задачи о вершинном покрытии графа к задаче отбора минимального по мощности множества эталонов.

Покажем связь этой задачи с задачей минимизации модифицированного функционала полного скользящего контроля $Q_k^*(\mu_\Omega)$.

В данной главе рассматриваются задачи классификации на два класса. Для удобства изложения и визуализации классы названы классом белых и черных объектов.

Теорема 2.2.1. *Задача поиска минимального размера вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω минимальной мощности, по которому классификация алгоритмом ближайшего соседа γ_Ω даст $\nu(\gamma_\Omega, X_G^L \backslash \Omega) = 0$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.*

Доказательство. По заданному графу G построим выборку X_G^L следующим образом.

Пусть в X_G^L будет два класса: черные объекты и белые объекты. Каждой вершине графа и каждому ребру графа поставим в соответствие по одному черному объекту, и добавим один белый объект, равноудаленный от всех черных.

Зададим попарные расстояния между всеми объектами. Между двумя черными объектами назначим расстояние r в следующих случаях: 1) когда эти два объекта соответствуют двум вершинам, соединенным в G ребром; 2) когда один из объектов соответствует ребру графа G, а второй соответствует вершине графа G, являющейся одним из концов этого ребра. Зададим расстояния от черных объектов до белого одинаковыми и равными $r_1 > r$. Все неоговоренные выше расстояния положим равными R, $R > r_1$.

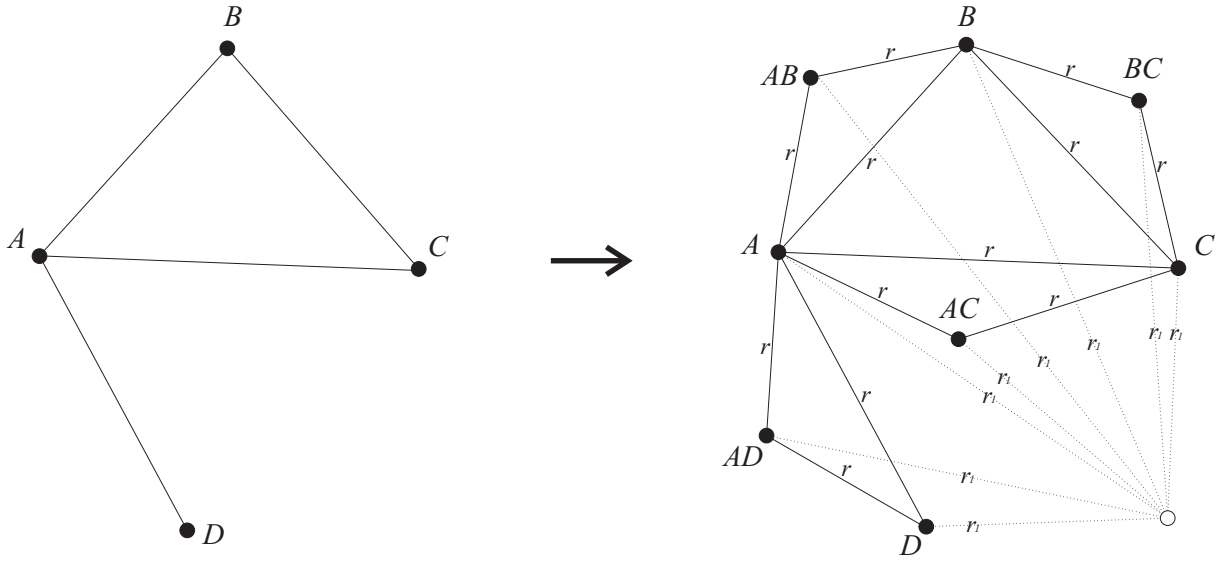


Рисунок 2.1 — Пример графа G, состоящего из четырёх вершин, и построение искусственной выборки X_G^L для случая модифицированного функционала полного скользящего контроля $Q_k^*(\mu_\Omega)$.

Заметим, что для выполнения неравенства треугольника в построенной выборке X_G^L, достаточно задать расстояния так, чтобы $2r > R > r_1 > r$, что легко обеспечить.

Таким образом, выборка X_G^L построена, см. пример на Рис. 1.

Пусть из выборки X_G^L выделено подмножество Ω такое, что $\nu(\gamma_\Omega, X_G^L \backslash \Omega) = 0$. Для доказательства теоремы достаточно показать, что Ω является вершинным покрытием исходного графа G.

Белый объект точно принадлежит Ω, иначе на нем была бы ошибка. Далее рассмотрим черные объекты.

Рассмотрим произвольную тройку объектов, соответствующих ребру AB и его вершинам A, B. Хотя бы один из этих объектов принадлежит Ω, иначе ближайшим эталонным к объекту AB был бы белый, и частота ошибок была бы не равна нулю.

Если в треугольнике A, B, AB эталонными являются все три объекта, то при удалении объекта AB из эталонных частота ошибок останется равной нулю, что противоречит минимальности Ω. Значит, что если Ω имеет минимальный размер, то не существует тройки объектов, соответствующих ребру и его вершинам, которая вся лежит в Ω.

Теперь предположим, что для минимального Ω существует тройка объектов A, B, AB такая, что $AB \in \Omega$. Тогда, если мы заменим AB в Ω на A или

B, который раньше Ω не принадлежал, то частота ошибок останется равной нулю, а мощность Ω не увеличится. Действительно, добавление объекта, соответствующего вершине, не может увеличить частоту ошибок, т. к. на белый он не повлияет, а для остальных черных ближайший сосед так и останется черным. Удаление объекта AB на объекты A и B не повлияет, т. к. один из них будет в Ω, следовательно, ближайшим эталонным для них будет черный. Для остальных черных он будет лежать дальше белого, который принадлежит Ω, значит его удаление не повлияет на классификацию остальных черных.

Итак, мы показали, что если построено минимальное Ω такое, что $\nu(\gamma_\Omega, X_G^L \backslash \Omega) = 0$, то можно построить Ω_1 такой же мощности, также обладающее свойством $\nu(\gamma_{\Omega_1}, X_G^L \backslash \Omega_1) = 0$, но при этом все черные объекты, лежащие в Ω_1, соответствуют только вершинам (но не ребрам) исходного графа G. Заметим, что при этом для каждой тройки объектов A, B, AB, соответствующих ребру AB и его вершинам A, B, хотя бы один из объектов A, B является эталонным.

Отметим в графе G все те вершины, у которых есть соответствующие объекты в Ω_1. То, что отмеченные вершины образуют покрытие, следует из того, что в X_G^L для каждой тройки объектов вида A, B, AB хотя бы один из объектов является эталонным, а среди Ω_1 все черные объекты соответствуют вершинам. Значит, для каждого ребра графа G хотя бы одна из его вершин отмечена.

Итак, мы доказали, что если единственный белый объект принадлежит Ω_1, то необходимо, чтобы вершины, соответствующие объектам Ω_1 являлись вершинным покрытием. Покажем, что это является достаточным условием для того, чтобы $\nu(\gamma_{\Omega_1}, X_G^L \backslash \Omega_1) = 0$.

Действительно, белый объект принадлежит эталонным, значит, алгоритм на нем не ошибается. Если x — один из объектов произвольной тройки A, B, AB, то либо сам x, либо его сосед на расстоянии r принадлежит Ω_1. Значит, есть черный эталонный объект ближе белого, следовательно, алгоритм на нем не ошибается, и частота ошибок равна нулю.

Минимальность полученного вершинного покрытия следует из минимальности Ω_1. Белый объект в Ω_1 должен быть добавлен обязательно, а если бы существовало вершинное покрытие меньшего размера, то можно было бы построить Ω_2, все еще дающее $\nu(\gamma_{\Omega_2}, X_G^L \backslash \Omega_2) = 0$, но при этом с меньшим

числом черных объектов. Для этого просто следует добавить в Ω_2 белый и все черные, соответствующие вершинному покрытию G. Получили бы противоречие минимальности Ω_1.

Итак, мы свели задачу поиска минимального размера вершинного покрытия к задаче поиска множества эталонных объектов Ω минимального размера, по которому классификатор ближайшего соседа γ_Ω даст $\nu(\gamma_\Omega, X_G^L \backslash \Omega) = 0$.

При этом размер X_G^L равен $1 + |V| + |E| = O(|V| + |E|)$, а построение выборки и преобразование эталонного множества Ω_1 в вершинное покрытие минимального размера осуществляется за линейное по $|V| + |E|$ число шагов.

Что и требовалось доказать.

Следствие 2.2.0.1. *Задача о поиске эталонного множества Ω размера, не превышающего h и дающего $\nu(\gamma_\Omega, X^L \backslash \Omega) = 0$, является NP-трудной.*

Заметим, что задача поиска множества эталонных объектов Ω, без ограничения по размеру, и дающего $\nu(\gamma_\Omega, X^L) = 0$, не является NP-трудной, поскольку имеет тривиальное решение $\Omega = X^L$.

Теперь рассмотрим связь задачи о минимальном вершинном покрытии графа с задачей минимизации функционала полного скользящего контроля по эталонному множеству Ω.

Сначала отдельно рассмотрим случай, когда длина контрольной выборки $k = 1$.

Теорема 2.2.2. *Задача поиска минимального размера вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω, минимизирующего функционал $Q_1(\mu_\Omega)$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.*

Предполагается, что в Ω входит хотя бы по одному объекту из каждого класса.

Доказательство. По заданному графу G построим выборку X_G^L следующим образом.

Пусть в X_G^L будет два класса: черные объекты и белые объекты. Каждой вершине A графа поставим в соответствие пару черных объектов A, A'.

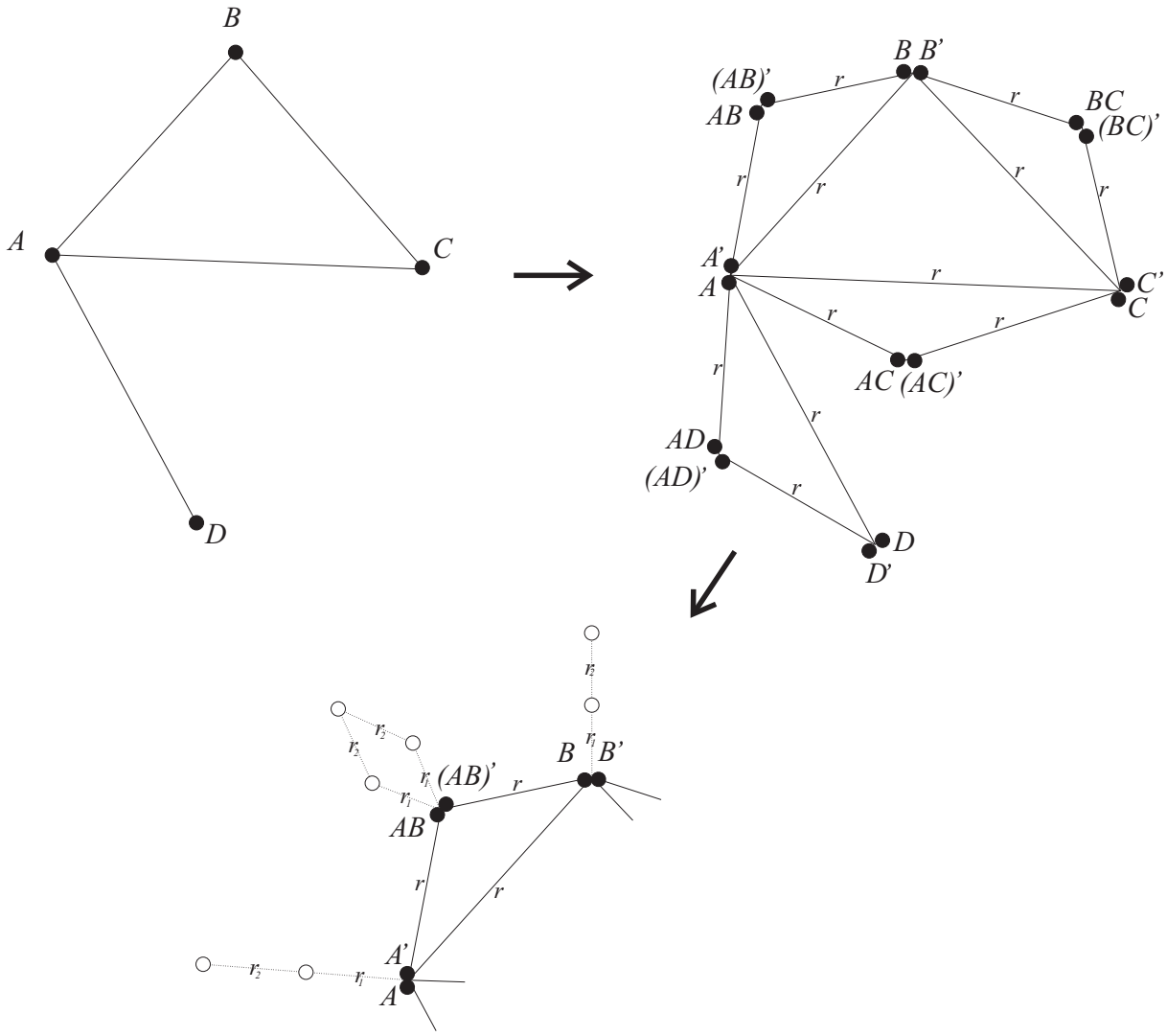


Рисунок 2.2 — Пример графа G, состоящего из четырёх вершин, и построение искусственной выборки X_G^L для случая функционала полного скользящего контроля $Q_1(\mu_\Omega)$.

Каждому ребру AB графа поставим в соответствие пару черных объектов AB, $(AB)'$.

Расстояния между объектами одной пары положим равными ε. Все остальные расстояния между объектами вида A, A', B, B', AB, $(AB)'$ положим равными $r > \varepsilon$.

Для каждой пары объектов, соответствующих ребру, добавим два белых объекта на расстоянии $r_1 > r$ от каждого из объектов пары, и один на расстоянии $r_2 > r_1$ от этих двух белых объектов.

Для каждой пары объектов, соответствующих вершине, добавим один белый объект на расстоянии $r_1 > r$ до каждого из объектов пары, и один на расстоянии $r_2 > r_1$ от этого белого объекта.

Все неоговоренные расстояния между парами черный-белый объект положим равными $r_3 > r_2$, а между парами объектов одинакового цвета — $R > r_3$.

Заметим, что справедлива цепочка неравенств $R > r_3 > r_2 > r_1 > r > \varepsilon$, которая, при дополнительном предположении $2\varepsilon > R$, обеспечивает выполнение неравенства треугольника.

Итак, выборка X_G^L построена, см. пример на Рис. 2.

Предположим, что множество эталонных объектов Ω минимизирует функционал $Q_1(\mu_\Omega)$. Для доказательства теоремы достаточно показать, что Ω соответствует вершинному покрытию исходного графа G.

Рассмотрим объекты A, A', B, B', AB, $(AB)'$, и все достроенные к ним белые объекты.

Возможны четыре случая.

1. Ни один из этих шести черных объектов не принадлежит Ω. Покажем что это противоречит минимальности $Q_1(\mu_\Omega)$.

В этом случае для пары объектов AB, $(AB)'$ не будет ни одного эталонного черного объекта, лежащего на расстоянии ближе, чем R, и принадлежащего к эталонным. Значит, ближайший к ним эталонный будет лежать либо на расстоянии r_3, либо на расстоянии r_1 и будет белым. Значит, ошибка на рассматриваемом множестве объектов будет не менее $2C(1)$. Добавим в Ω объекты A, A' и семь белых объектов, достроенных к рассматриваемым трем парам. На остальном множестве объектов ошибка не увеличится, так как добавленные в Ω черные объекты A, A' не могут увеличить ошибку на черных, а на белые, кроме указанных семи, никак не влияют, поскольку рассматривается один сосед ($k = 1$), а черные объекты на таком же расстоянии от них уже есть. Добавленные в Ω белые объекты не увеличат ошибку на остальной выборке, так как для белых они увеличить ошибку не могут, а, поскольку рассматривается один сосед ($k = 1$), для всех черных объектов вне рассматриваемых трех пар белые на таком расстоянии уже есть.

Рассмотрим ошибку на A, A', B, B', AB, $(AB)'$ и семи белых объектах, достроенных к ним. Для всех этих черных объектов ошибка будет равна нулю, так как ближайший сосед из Ω будет находится на расстоянии либо ε, либо r, и будет черным. Белые же эталонные объекты будут находится на расстоянии не менее r_1. Для пяти белых объектов, достроенных к парам черных, в которых нет эталонных, ближайший сосед из эталонных будет на расстоянии r_2, и будет

белым. А ближайший черный из Ω будет на расстоянии $r_3 > r_2$. Значит, на этих пяти объектах ошибки нет.

Рассмотрим два оставшихся белых объекта, достроенных к паре черных, добавленных в Ω. Ошибка на них в сумме будет равняться $C(1)$.

Итак, $2C(1) > C(1)$, следовательно, такая операция над Ω улучшила бы функционал $Q_1(\mu_\Omega)$, что противоречит предположению о его минимальности. Значит, если Ω минимизирует функционал $Q_1(\mu_\Omega)$, то среди каждой тройки групп, соответствующих ребру графа G и двум вершинам, между которыми это ребро проведено, есть хотя бы один объект, принадлежащий Ω, что и требовалось показать.

2. Пусть хотя бы один из объектов AB, $(AB)'$ принадлежит Ω, а объекты A, A', B, B' не принадлежат Ω. Посчитаем ошибку на достроенных к ним семи белых объектах. Для двух объектов, находящихся на расстоянии r_1 от AB, $(AB)'$ ближайшим эталонным будет черный из этой группы. Значит, ошибка на шести рассматриваемых черных и семи достроенных к ним белых объектах будет не менее $2C(1)$.

Удалим из Ω эталонные объекты пары AB, $(AB)'$ и добавим в Ω все семь рассматриваемых белых объектов и черные A, A'. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится, при этом ошибка на рассматриваемых шести черных объектах и семи белых станет $C(1)$. Значит, $Q_1(\mu_\Omega)$ уменьшится, что противоречит предположению о его минимальности. Следовательно, случай 2 для Ω, минимизирующего $Q_1(\mu_\Omega)$, реализоваться не может.

3. Пусть среди пары (A, A') и пары $(AB, (AB)')$ есть объекты в Ω. Посчитаем ошибку на достроенных семи объектах. Для двух объектов, находящихся на расстоянии r_1 от пары черных AB, $(AB)'$, ближайшим эталонным будет черный из этой группы. Для белого объекта, находящегося на расстоянии r_1 от A, A', один из этих черных и будет его ближайшим соседом из эталонных. Значит, ошибка на шести рассматриваемых черных и достроенных к ним белых объектах будет не менее $3C(2)$.

Удалим из Ω эталонные объекты пары AB, $(AB)'$, добавим в Ω все семь рассматриваемых белых объектов и все черные A, A', B, B'. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится, а на рассматриваемых шести черных и семи белых объектах станет $2C(1)$. Значит, $Q_1(\mu_\Omega)$

уменьшится, что противоречит предположению о минимальности $Q_1(\mu_\Omega)$. Значит, случай 3 для Ω, минимизирующего $Q_1(\mu_\Omega)$, реализоваться не может.

4. Пусть в каждой из пар (A, A'), (B, B'), $(AB, (AB)')$ есть объекты, принадлежащие Ω. Посчитаем ошибку на семи достроенных к ним белых объектах. Для двух объектов, находящихся на расстоянии r_1 от пары черных AB, $(AB)'$, ближайшим эталонным будет черный из этой пары. Для белого объекта, находящегося на расстоянии r_1 от пары B, B', один из них и будет ближайшими эталонным объектом к данному белому. Аналогично для белых, достроенных к A, A'. Значит, ошибка на шести рассматриваемых черных объектах и достроенных к ним семи белых будет не менее $4C(2)$.

Удалим из Ω эталонные объекты пары AB, $(AB)'$, добавим в Ω все семь рассматриваемых белых объекта и все черные A, A', B, B'. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится, а на рассматриваемых шести черных и семи белых объектах станет $2C(1)$. Значит, $Q_1(\mu_\Omega)$ уменьшится, что противоречит предположению о минимальности $Q_1(\mu_\Omega)$. Значит, случай 4 для Ω, минимизирующего $Q_1(\mu_\Omega)$ реализоваться не может.

Отметим в графе G все вершины, которым соответствует хотя бы один эталонный объект в Ω. Если Ω минимизирует $Q_1(\mu_\Omega)$, то в Ω нет объектов, соответствующих ребрам G. С другой стороны, для каждого ребра графа G среди объектов X^L_G, соответствующих его вершинам, есть эталонные. Отсюда следует, что отмеченные вершины образуют вершинное покрытие графа G.

Покажем теперь минимальность этого покрытия.

Если среди Ω есть эталонные объекты, соответствующие ровно m различным вершинам, то $Q_1(\mu_\Omega) = mC(1)$, так как меньше $mC(1)$ быть не может из-за ошибок на белых, достроенных к группам черных объектов, соответствующих вершинам; при этом случай $mC(1)$ можно реализовать, добавив в Ω все белые объекты, и все такие черные, которые соответствуют этим выше указанным m вершинам.

Если бы для графа G существовало вершинное покрытие размера m^*, меньшего m, то по нему можно было бы построить Ω^*, дающее $Q_1(\mu_{\Omega^*}) = m^*C(1)$, просто назначив эталонными все белые объекты, и все объекты, соответствующие вершинам, принадлежащим минимальному вершинному покрытию. Это противоречит предположению о том, что Ω минимизирует $Q_1(\mu_\Omega)$.

Итак, мы показали способ построения минимального вершинного покрытия по известному Ω, то есть свели задачу поиска минимального вершинного покрытия произвольного графа G к задаче выбора из некоторой искусственной выборки X_G^L минимального по мощности множества эталонных объектов, минимизирующего функционал $Q_1(\mu_\Omega)$. Причем и время построения X_G^L, и время преобразования множества Ω в вершинное покрытие, и размер выборки, линейны по $|V| + |E|$.

Теорема доказана.

Теперь рассмотрим случай $k \geq 2$.

Теорема 2.2.3. *Задача поиска минимального вершинного покрытия произвольного графа G сводится к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов Ω минимальной мощности, минимизирующего функционал $Q_k(\mu_\Omega)$. Причем выборка X_G^L строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.*

Предполагается, что в Ω содержится не менее $k+1$ объектов каждого класса[1].

Доказательство. По заданному графу G построим выборку X_G^L следующим образом.

Пусть в X_G^L будет два класса: черные объекты и белые объекты. Пусть AB — произвольное ребро графа с вершинами A и B. Каждой вершине A графа поставим в соответствие группу из $k+1$ черных объектов $\{A_i\} = \{A_i : i = 1, \ldots, k+1\}$. Каждому ребру AB графа поставим в соответствие группу из $k+1$ черных объектов $\{(AB)_i\} = \{(AB)_i : i = 1, \ldots, k+1\}$. Расстояния между объектами внутри одной группы положим равными ε. Расстояния между объектами из разных групп $\{A_i\}$, $\{B_i\}$ и $\{(AB)_i\}$ положим равными $r > \varepsilon$.

Для каждой группы объектов $\{(AB)_i\}$, соответствующих ребру, добавим пару белых объектов на расстоянии $r_1 > r$ от каждого из объектов группы, и один отдельный на расстоянии $r_2 > r_1$ от этой пары белых объектов. Еще $(k-1)$ белых объекта с попарными расстояниями ε добавим на расстоянии r от этого отдельного белого и расстоянии ε до пары белых, указанной выше.

[1] При помощи дополнительных построений далее будет показано, что данное предположение делается без ограничения общности.

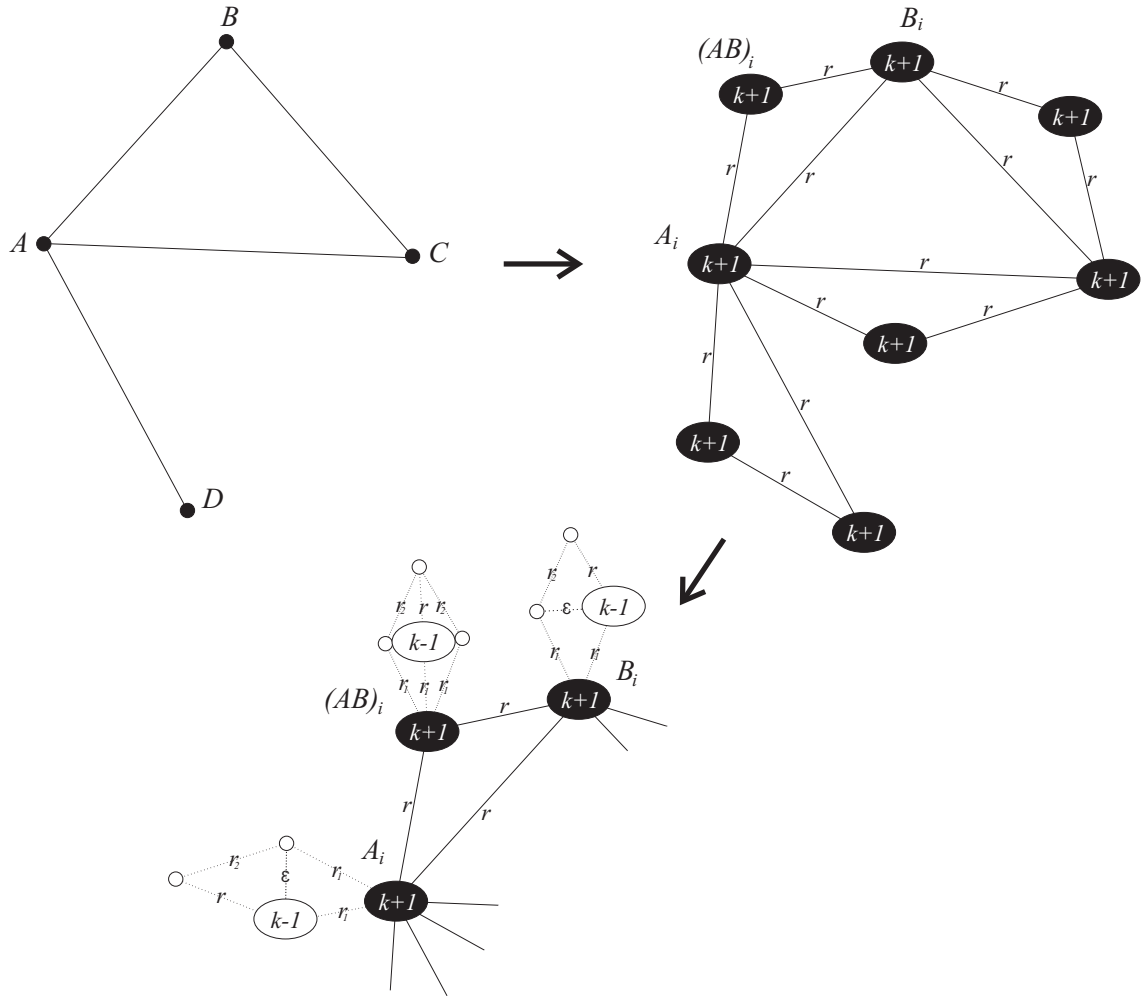


Рисунок 2.3 — Пример графа G, состоящего из четырёх вершин, и построение искусственной выборки X_G^L для случая функционала полного скользящего контроля $Q_k(\mu_\Omega)$, $k \geq 2$.

Расстояние от этой группы из $(k-1)$ белых объектов до рассматриваемой группы черных положим равным r_1.

Для каждой группы объектов $\{A_i\}$, соответствующей вершине, добавим один белый объект на расстоянии $r_1 > r$ до каждого из объектов группы, один отдельный белый на расстоянии $r_2 > r_1$ от этого белого объекта. Добавим еще группу из $k-1$ белых объектов с попарными расстояниями ε так, чтобы их расстояние до белого, находящегося от черных на расстоянии r_1, составляло ε, а расстояние до отдельного белого положим равным r. Расстояние от этой группы[2] из $(k-1)$ белых объектов до рассматриваемой группы черных положим равным r_1.

[2]Под расстоянием между двумя группами объектов понимается расстояние межу каждой парой объектов, взятых по одному из этих групп.

Все неоговоренные расстояния между парами объектов черный–белый положим равными $r_3 > r_2$, а между парами объектов одинакового цвета — $R > r_3$.

Заметим, что справедлива цепочка неравенств $R > r_3 > r_2 > r_1 > r > \varepsilon$, которая, при дополнительном предположении $2\varepsilon > R$, обеспечивает выполнение неравенства треугольника.

Итак, выборка X_G^L построена, см. пример на Рис. 3.

Предположим, что множество эталонных объектов Ω минимизирует функционал $Q_k(\mu_\Omega)$. Для доказательства теоремы достаточно показать, что Ω соответствует вершинному покрытию исходного графа G.

Рассмотрим тройки групп $\{A_i\}$, $\{B_i\}$, $\{(AB)_i\}$, соответствующих произвольному ребру AB графа G, двум его вершинам A, B, между которыми это ребро проведено, и все достроенные к ним белые объекты.

Рассмотрим четыре случая:

1. Ни один из черных объектов групп $\{A_i\}$, $\{B_i\}$, $\{(AB)_i\}$ не принадлежит Ω. Но тогда для группы объектов, соответствующих ребру, не будет ни одного черного объекта, лежащего на расстоянии ближе, чем R, и принадлежащего множеству эталонов Ω. Значит, ближайший к ним эталонный объект будет лежать либо на расстоянии r_3, либо на расстоянии r_1, и будет белым. Следовательно, ошибка на рассматриваемом множестве объектов составит не менее $(k+1)C(1)$. Добавим в эталонные объекты все объекты группы $\{A_i\}$ и все белые объекты, достроенные к рассматриваемым трем группам $\{A_i\}$, $\{B_i\}$, $\{(AB)_i\}$. На остальном множестве объектов ошибка не увеличится, поскольку добавленные в эталонные черные объекты не могут увеличить ошибку на черных, а на белые, не достроенные к $\{A_i\}$, $\{B_i\}$, $\{(AB)_i\}$, никак не влияют, поскольку k черных объектов на таком же расстоянии от них уже есть. Добавленные в Ω белые объекты не увеличат ошибку на остальной выборке, поскольку для белых они увеличить ошибку не могут, а для всех черных объектов вне рассматриваемых трех групп k белых на таком расстоянии уже есть.

Итак, осталось рассмотреть поведение ошибки на трех рассматриваемых группах черных и достроенных к ним белых объектах. Для всех рассматриваемых черных объектов ошибка будет равна нулю, так как k ближайших соседей из Ω будут находится на расстоянии либо ε, либо r, и будут черными. Белые же эталонные объекты будут находится на расстоянии не менее r_1. Для белых

объектов, достроенных к группам черных, в которых нет эталонных, k ближайших соседей из эталонных будут находиться на расстоянии r_2 или ε, и будут белыми. А ближайший черный из Ω будет на расстоянии $r_3 > r_2$. Значит, на этих объектах ошибки нет.

Рассмотрим оставшиеся белые объекты, достроенные к группе черных, добавленных в эталонные. Ошибка на них в сумме будет равна $C(k)$.

Итак, $2C(1) > C(k)$, значит, такая операция над Ω уменьшила бы функционал $Q_k(\mu_\Omega)$, что противоречит предположению о его минимальности. Значит, если Ω минимизирует функционал $Q_k(\mu_\Omega)$, то среди каждой тройки групп, соответствующих ребру графа G и двум вершинам, между которыми это ребро проведено, есть хотя бы один объект, принадлежащий Ω.

2. Пусть среди объектов групп $\{A_i\}$, $\{B_i\}$ нет эталонных объектов, но они есть в группе $\{(AB)_i\}$. Посчитаем ошибку на достроенных к этим трем группам белых объектах. Для двух объектов, находящихся на расстоянии r_1 от группы черных $\{(AB)_i\}$, одним из k ближайших эталонных будет черный из этой группы. Значит, ошибка на трех рассматриваемых группах и достроенных к ним белых будет не менее $2C(k)$.

Удалим из Ω эталонные объекты, соответствующие ребру AB, добавим в Ω все рассматриваемые белые и все черные объекты группы $\{A_i\}$. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится, а ошибка на рассматриваемых трех группах черных объектов и достроенных к ним белых объектах станет равна $C(k)$. Значит, $Q_k(\mu_\Omega)$ уменьшается, что противоречит предположению о его минимальности. Значит случай 2 для Ω, минимизирующего $Q_k(\mu_\Omega)$, реализоваться не может.

3. Пусть среди объектов трех рассматриваемых групп есть эталонные либо среди $\{A_i\}$, либо среди $\{(AB)_i\}$. Посчитаем ошибку на достроенных к этим трем группам белых объектах. Для двух объектов, находящихся на расстоянии r_1 от группы черных $\{(AB)_i\}$, одним из ближайших k эталонных будет черный из этой группы. Для белого объекта, находящегося на расстоянии r_1 от группы $\{A_i\}$, эталонный черный объект из $\{A_i\}$ будет одним из k ближайших эталонных объектов. Значит, ошибка на трех рассматриваемых группах черных и достроенных к ним белых объектах будет не менее $3C(k)$.

Удалим из Ω эталонные объекты, соответствующие этому ребру, добавим в Ω все рассматриваемые белые и все из групп черных, соответствующих вер-

шинам. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится. На рассматриваемых трех группах черных объектов и достроенных к ним белых объектах ошибка станет $2C(k)$. Значит, $Q_k(\mu_\Omega)$ уменьшится, что противоречит предположению о его минимальности. Значит, случай 3 для Ω, минимизирующего $Q_k(\mu_\Omega)$, реализоваться не может.

4. Пусть среди объектов трех рассматриваемых групп во всех трех группах есть объекты, принадлежащие Ω. Посчитаем ошибку на достроенных к этим трем группам белых объектах. Для двух объектов, находящихся на расстоянии r_1 от группы черных $\{(AB)_i\}$, одним из k ближайших эталонных будет черный объект из этой группы. Для белого объекта, находящегося на расстоянии r_1 от группы черных, соответствующих вершине, эти эталонные черные объекты будут одними из k ближайших эталонных объектов. Значит, ошибка на трех рассматриваемых группах и достроенных к ним белых будет не менее $4C(k)$.

Удалим из Ω эталонные объекты, соответствующие этому ребру, добавим в Ω все рассматриваемые белые и все из групп черных, соответствующих вершинам. По причинам, описанным в случае 1, ошибка на остальной выборке не увеличится. При этом ошибка на рассматриваемых трех группах черных объектов и достроенных к ним белых станет равна $2C(k)$. Значит, $Q_k(\mu_\Omega)$ уменьшится, что противоречит предположению о его минимальности. Значит, случай 4 для Ω, минимизирующего $Q_k(\mu_\Omega)$, реализоваться не может.

Отметим для G все вершины, которым соответствует хотя бы один эталонный объект в Ω. Итак, мы получили, что если Ω минимизирует $Q_k(\mu_\Omega)$, то в Ω нет объектов, соответствующих ребрам G. С другой стороны, для каждого ребра графа G среди объектов X^L_G, соответствующих его вершинам, есть эталонные.

Отсюда следует, что отмеченные вершины образуют вершинное покрытие графа G. Покажем теперь минимальность этого покрытия.

Если среди Ω есть эталонные объекты, соответствующие ровно m различным вершинам, то $Q_k(\mu_\Omega) = kC(1)$, так как меньше $mC(1)$ быть не может из-за ошибки на белых, достроенных к группам черных объектов, соответствующих вершинам, а случай $mC(1)$ можно реализовать, добавив в Ω все белые объекты и все такие черные, которые соответствуют указанным m вершинам.

Если бы для графа G существовало вершинное покрытие размера m^*, меньшего m, то по нему можно было бы построить Ω^*, дающее $Q_k(\mu_{\Omega^*}) = m^*C(1)$, просто назначив эталонными все белые объекты и все объекты, соответствующие вершинам, принадлежащим минимальному вершинному покрытию. Что противоречит предположению о том, что Ω минимизирует $Q_k(\mu_\Omega)$.

Итак, мы показали способ построения минимального вершинного покрытия по известному Ω, то есть свели задачу поиска минимального по мощности вершинного покрытия произвольного графа G к задаче выбора из некоторой искусственной выборки X_G^L множества эталонных объектов, минимизирующего функционал $Q_k(\mu_\Omega)$, при $k \geq 1$. Причем и время построения X_G^L, и преобразование Ω в вершинное покрытие, и размер выборки имеют порядок (has order) $O(k|E| + k|V|)$.

Теорема доказана.

Приведенное выше доказательство предполагало, что в Ω лежит не менее $k+1$ объектов каждого из классов.

Теорема 2.2.4. *В условии теоремы 2.2.3 предположение о том, что в Ω лежит не менее $k+1$ объектов каждого класса, можно заменить предположением $|\Omega| \geq k+1$.*

Доказательство. Дополним уже описанную выше выборку X_G^L множеством объектов, обеспечивающих требуемые $k+1$ объектов каждого класса среди Ω, минимизирующего $Q_k(\mu_\Omega)$, при условии $|\Omega| \geq k+1$. Добавим по $k+1$ групп черных и белых объектов, состоящих их $k+1$ объектов одинакового цвета в каждой группе. Объекты внутри группы находятся на расстоянии ε, а неоговоренные расстояния строятся по тем же правилам, что и ранее: между парами объектов одного цвета R, а разных r_3. Количество объектов, которыми мы дополнили X_G^L, составляет $O(k^2)$, что не нарушает полиномиальности длины выборки X_G^L по L. Обозначим дополненную выборку через $X_G^{L'}$, где L' — длина дополненной выборки.

Рассмотрим Ω^*, минимизирующее функционал полного скользящего контроля $Q_k(\mu_\Omega)$ в этом случае. Предполагается, что в Ω^* есть хотя бы по одному объекту из каждого класса и не менее $k+1$ объектов всего. Разделим последовательность соседей каждого из объектов дополненной выборки $X_G^{L'}$ на три подпоследовательности, в порядке возрастания расстояний:

1) объекты, расстояние до которых меньше r_3;

2) объекты чужого класса, находящиеся на расстоянии r_3;

3) объекты своего класса, находящиеся на расстоянии R.

Рассмотрим любую из одноцветных групп $X_G^{L'} \setminus X_G^L$, в которой нет эталонного объекта. Для каждого объекта из этой группы ближайший к нему эталонный объект принадлежит другому классу. Если мы добавим один из объектов рассматриваемой группы в эталонные, то для каждого из оставшихся k объектов этой группы он станет ближайшим эталоном, сдвинув последовательность ближайших объектов на 1 вправо; при этом объект, который был ближайшим соседом, станет вторым соседом. Значит, на рассматриваемой одноцветной группе функционал уменьшится не менее, чем на $kC(1) - kC(2)$. Для объекта, который мы добавили в эталонные, последовательность его соседей не изменится. Для остальных объектов того же цвета, но вне этой группы, ошибка не увеличится. Для объектов противоположного цвета ошибка увеличится не более, чем на $NC(2)$, где N — количество объектов противоположного цвета в $X_G^{L'}$. Добавляемый объект будет лежать в подпоследовательности 2) для каждого из объектов противоположного цвета. Значит, для каждого из объектов противоположного цвета ошибка, связанная с увеличением порядкового номера последующих соседей, увеличится не более чем на $C(2)$.

Следовательно, функционал полного скользящего контроля уменьшится не менее чем на $kC(1) - (k+N)C(2)$. Из определения коэффициентов $C(m)$ следует $\frac{C(1)}{C(2)} = \frac{L'-2}{k-1}$. Поскольку $k \geq 2$, и число объектов каждого из цветов не менее $k^2 \geq 2k$ по построению, имеет место цепочка неравенств: $N + k \leq L' - k \leq L' - 2$. Тогда

$$\frac{C(1)}{C(2)} = \frac{L'-2}{k-1} \geq \frac{N+k}{k-1} > \frac{N+k}{k},$$

следовательно, $kC(1) - (k+N)C(2) > 0$. Это означает, что после добавления рассматриваемого эталонного объекта функционал полного скользящего контроля уменьшится, что противоречит предположению о его минимальности. Значит, в каждой группе из $X_G^{L'} \setminus X_G^L$ есть хотя бы один эталонный объект. Поэтому число эталонных объектов каждого цвета — не менее $k+1$. Заметим, что если все объекты $X_G^{L'} \setminus X_G^L$ добавить в эталонные, то ошибка на остальных объектах не изменится, а на них станет равной нулю. Теперь к выборке $X_G^{L'}$

можно применить все те же рассуждения, что и при доказательстве теоремы 2.2.3, однако теперь от Ω требуется только $|\Omega| \geq k+1$.

Теорема доказана.

Следствие 2.2.0.2. *Для любого k задача поиска множества эталонных объектов Ω, $|\Omega| \geq k+1$, для которого функционал $Q_k(\mu_\Omega)$ не превышает h, является NP-трудной.*

Заметим, что условие $|\Omega| \geq k+1$ действительно необходимо, иначе функционал Q невозможно будет записать в виде, указанном в теореме 3.3.1.

NP-трудность этих задач обосновывает применение различных эвристических алгоритмов, выбирающих Ω так, что минимизируемый функционал принимает значение не наименьшее, но близкое к наименьшему, и/или мощность множества Ω не минимальна, но близка к минимальной.

2.3 Вычислительная сложность задачи отбора признаков

Пусть дана обучающая выборка X^L, описываемая множеством признаков $\mathbb{F}$. Компактность выборки зависит от функции расстояния ρ, при помощи которой измеряется сходство. Для оценки компактности выборки будем использовать первый профиль компактности $P(1)$, который совпадает со значением функционала Q_1 для алгоритма ближайшего соседа.

Функция ρ обычно строится по признаковым описаниям объектов. Одними из самых используемых функций ρ является взвешенное расстояние Минковского.

$$\rho_{p\text{Mink}}(x, x') = \Big(\sum_{d=1}^{|\mathbb{F}|} w_d |f_d(x) - f_d(x')|^p\Big)^{\frac{1}{p}},$$

где $f_d(x)$ — значение признака f_d для объекта x, а w_d — вес признака f_d, $w_d \geq 0$, $\sum_{d=1}^{|\mathbb{F}|} w_d = 1$.

Иногда в прикладных задачах признаки $f_d \in \mathbb{F}$ задаются не значениями для каждого объекта $f_d(x)$, а матрицами попарных расстояний $\rho_d(x,x')$ по данному признаку. В этом случае расстояние Минковского можно выразить

следующим образом:

$$\rho_{p\text{Mink}}(x, x') = \Big(\sum_{d=1}^{n} w_d \rho_d^p(x,x')\Big)^{\frac{1}{p}}.$$

Даже если признаки изначально заданны значениями для каждого из объектов, то при подсчете $\rho_{p\text{Mink}}$ для каждой пары объектов вычисляются попарные расстояния по каждому из признаков, то есть вычисляется соответствующая матрица попарных расстояний.

Вклад каждого объекта x в $P(m)$ зависит не от конкретных значений функции расстояний ρ, а от того, как упорядочена остальная выборка с точки зрения возрастания ρ до объекта x. Поскольку возведение в степень является монотонным преобразованием, значения $P(m)$ для функций расстояния $\rho = \rho_{p\text{Mink}}$ и $\rho = \rho^p_{p\text{Mink}}$ совпадают.

Тогда оценку компактности выборки с функцией расстояний $\rho_{p\text{Mink}}$ при помощи функции $P(1)$ можно свести к следующему частному случаю: признаки $f \in \mathbb{F}$ заданы матрицей попарных расстояний $\rho_f(x,x')$ между объектами, а возможные функции расстояний $\rho_{\mathbb{F}}(x,x')$ — линейными комбинациями $\rho_f(x,x')$, $f \in \mathbb{F}$ с неотрицательными весами. Для сведения задачи с $\rho_{p\text{Mink}}$ к задаче с $\rho_{\mathbb{F}}$ достаточно достаточно каждый элемент матриц попарных расстояний возвести в степень p.

Любая функция расстояний ρ_F, построенная на подмножестве признаков $F \subseteq \mathbb{F}$, может быть построена и на $\mathbb{F}$. Следовательно, множество признаков $\mathbb{F}$ всегда не хуже любого своего подмножества $F \subseteq \mathbb{F}$ с точки зрения $P(1)$. Однако чем больше признаков мы используем, тем сильнее модель склонна к переобучению.

Опр. 2.3.1. *Минимальное значение профиля компактности $P(1)$, достижимое при использовании функций расстояний, построенных на множестве признаков F, будем называть минимумом $P(1)$ на F.*

Поставим задачу отбора признаков следующим образом:

Задача 2.3.0.1. *Выбрать минимальное по количеству множество признаков $F \subseteq \mathbb{F}$ такое, что минимум $P(1)$ на $\mathbb{F}$ совпадает с минимумом $P(1)$ на F.*

Теорема 2.3.1. *Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 2.3.0.1 на некоторой искусственной обучающей выборке X^L.*

Доказательство. Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots C_{|S|}\}$ — семейство его подмножеств.

Пусть объекты X^L могут принадлежать одному из двух классов $\mathbb{A}$ и $\mathbb{B}$. Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Добавим два вспомогательных объекта a' и a^* класса $\mathbb{A}$ и один вспомогательный объект b класса $\mathbb{B}$. Каждому подмножеству C_t поставим в соответствие признак f_t. Таким образом X^L будет содержать $|U| + 3$ объекта, описываемых $|S|$ признаками.

Пусть каждый признак f_t соответствует следующим попарным расстояниям:

- расстояние между объектами a' и a^* равно r_0;
- расстояние между любыми двумя объектами a_i и a_j, $i \neq j$ равно R;
- расстояние от любого объекта класса $\mathbb{A}$ до b равно r_1;
- расстояние от любого объекта a_i до a^* равно R;
- расстояние от объекта a_i до a' равно r, если $u_i \in C_t$, и равно R в противном случае.

где $R > r_1 > r > r_0$; $R \cdot (|S| - 1) + r < r_1 \cdot |S|$, то есть $R - r > |S| \cdot (R - r_1)$

Рассмотрим подмножество признаков F. Какую бы функцию расстояний мы не взяли, для объекта b ближайший сосед всегда будет принадлежать другому классу и давать вклад $\frac{1}{|U|+3}$ в $P(1)$. Таким образом наименьшее значение $P(1)$ не менее $\frac{1}{|U|+3}$.

Объекты a' и a^* всегда будут ближайшими друг другу и давать нулевой вклад в $P(1)$.

Объект a_i может давать нулевой вклад в $P(1)$ только если хотябы по одному из признаков ближайшим для него является объект a', то есть если существует t такое, что $u_i \in C_t$ и $f_t \in F$. В противном случае ближайшим для объекта a_i будет b, и a_i будет давать вклад $\frac{1}{|U|+3}$ в $P(1)$.

Следовательно, чтобы минимальное значение $\frac{1}{|U|+3}$ функционала $P(1)$ достигалось, необходимо чтобы множество подмножеств, соответствующих F являлось покрытием множества U.

Предположим множество подмножеств $S_F \subseteq S$, соответствующих F является покрытием множества U. Пусть функция расстояний равна сумме всех признаков F с весами, равными $\frac{1}{|F|}$. Тогда расстояние от каждого объекта a_i до a' не больше чем $\frac{1}{|F|}(R \cdot (|F| - 1) + r)$, а до объекта b равно r_1. Поскольку $R - r > |S| \cdot (R - r_1) \geq |F| \cdot (R - r_1)$, следовательно $\frac{1}{|F|}(R \cdot (|F| - 1) + r) < r_1$, то есть ближайшим для объекта a_i будет объект того же класса. Получим $P(1) = \frac{1}{|U|+3}$.

Таким образом значение минимум функционала $P(1)$ равен $\frac{1}{|U|+3}$ и достигается на подмножестве признаков F тогда и только тогда, когда множество подмножеств $S_F \subseteq S$, соответствующих F является покрытием множества U. Минимальное множество признаков F соответствует $S_F \subseteq S$ минимальному покрытию множества подмножествами.

□

Следствие 2.3.0.1. *Задача выбора множества признаков $F \subseteq \mathbb{F}$ такого, что минимум $P(1)$ на $\mathbb{F}$ совпадает с минимумом $P(1)$ на F и $|F| \leq q$ является NP-трудной.*

2.4 Основные выводы главы 2

1. Из теоремы 2.1.1 следует, что функционал полного скользящего контроля для алгоритма ближайшего соседа может быть использован для оценки компактности выборки.
2. При условиях, описанных в 2.1, функционал полного скользящего контроля совпадает с частотой ошибок.
3. Отбор объектов с частотой ошибок в виде целевой функции без дополнительных ограничений приводит к вырожденным решениям. Если частота меряется только по отобранным объектам — оптимальным, но бесполезным с практической точки зрения является, например, решение, отбрасывающее все объекты одного из классов. Если частота меряется по всем объектам, а соседей рассматривают только среди отобранных — оптимальным будет решение, не отбрасывающее ни одного объекта, то есть отбор объектов осуществляться не будет. Сле-

довательно, для использования частоты ошибок в качестве целевой функции требуются дополнительные ограничения, запрещающие вырожденные решения.

4. Задачи отбора множества объектов, оптимального с точки зрения предложенных формализаций выполнения гипотезы компактности, являются NP-трудными.
5. Задача отбора множества признаков, оптимального с точки зрения предложенной формализации выполнения гипотезы компактности, является NP-трудной.
6. NP-трудность рассматриваемых задач отбора признаков и объектов обосновывает применение приближенных алгоритмов.

Глава 3. Отбор эталонов и признаков с ограничениями монотонности

Задачи отбора объектов и признаков являются важным этапом построения монотонных классификаторов. Монотонными называются классификаторы, для которых предполагается наличие монотонных зависимостей между значениями признаков и меткой классов.

Известны различные методы монотонной классификации [12; 14; 19–21]. Многие из них требуют монотонной обучающей выборки. Однако на практике данные, как правило, не идеальны, и монотонность на некоторых объектах нарушается. Существуют методы построения монотонного классификатора ближайшего соседа по немонотонной выборке [22], но без предобработки они слишком чувствительны к выбросам. Потому приходится производить так называемую *монотонизацию* обучающей выборки — искать подмножества объектов и признаков, для которых будут выполнятся ограничения монотонности. В некоторых подходах задача монотонизации производится при помощи изменения метки класса [28; 29].

В данной работе задача монотонизации рассматривается только как отбор объектов и признаков, без возможности изменения метки класса. Особенностью задач отбора объектов и признаков в данном случае является то, что кроме выполнения гипотезы компактности требуется еще и выполнение ограничений монотонности.

В данной главе устанавливается связь между метрическими и монотонными классификаторами, предлагается серия оптимизационных постановок задач отбора объектов и признаков с ограничениями монотонности, оценивается вычислительная сложность предложенных постановок.

Для упрощения изложения предполагается, что если для некоторого признака указывается только частичный порядок, то он может быть дополнен до линейного произвольным образом. Это возможно так как на конечном множестве всякий частичный порядок может быть дополнен до линейного.

3.1 Связь метрических и монотонных алгоритмов

Известна следующая связь между метрическими и монотонными классификаторами: использование монотонной функции расстояния, предложенной в работах [22–24] позволяет строить монотонный метрический классификатор. В данном разделе показано, что использование монотонной функции расстояний, упрощенной относительно работ [22–24], позволяет оценить точность работы монотонных классификаторов произвольного вида.

Рассмотрим понятие монотонности более формально.

Дано конечное множество $X^L = (x_i,y_i)_{i=1}^L$, называемое обучающей выборкой, и конечное множество признаков $\mathbb{F} = \{f_1,\ldots,f_t\}$ — отображений вида $f_j\colon X^L \to E_j$, где E_j — линейно упорядоченное множество. Каждый объект x_i относится к одному из двух классов $y_i = y^*(x_i) \in \{0,1\}$. Обозначим множество объектов обучающей выборки класса 1 через $\mathbb{A}$, а класса 0 — через $\mathbb{B}$.

Любое непустое подмножество множества признаков $F \subseteq \mathbb{F}$ индуцирует отношение частичного порядка на X^L: $x \le x'$ тогда и только тогда, когда $f(x) \le f(x')$ для всех $f \in F$; $x < x'$ тогда и только тогда, когда $x \le x'$ и $x \ne x'$.

Опр. 3.1.1. *Пара объектов $(a,b) \in \mathbb{A} \times \mathbb{B}$ называется монотонной, если $a > b$. Множество всех монотонных пар обозначается через M.*

Опр. 3.1.2. *Пара объектов $(a,b) \in \mathbb{A} \times \mathbb{B}$ называется дефектной, если $a < b$. Множество всех дефектных пар обозначается через D.*

Множество пар, монотонных по признаку f, будем обозначать через M_f, монотонных по совокупности признаков F — через M_F. Множество пар, дефектных по признаку f, будем обозначать через D_f, дефектных по совокупности признаков F — через D_F.

Замечание 3.1.1. *Из определения следует, что для монотонности по подмножеству признаков не требуется строгой монотонности по каждому из признаков. Однако для упрощения выкладок если не оговорено иначе будем считать, что значения признаков для объектов не совпадают.*

Опр. 3.1.3. *Выборка называется монотонной, если её объекты не образуют ни одной дефектной пары.*

Опр. 3.1.4. *Классификатор γ называется монотонным, если для любых двух объектов x, x' из $x < x'$ следует $\gamma(x) \leq \gamma(x')$. То есть классификатор монотонный, если для любой пары объектов меньший объект он относит к небольшему классу.*

Пусть дано некоторое монотонное эталонное множество Ω объектов с метками классов, которые монотонный алгоритм γ точно классифицирует. Тогда любые объекты, большие какого-то объекта класса 1 из множества Ω, а также объекты меньшие какого-то объекта класса 0 из множества Ω будут однозначно классифицироваться исходя из свойств монотонности. Классификация остальных объектов будет неоднозначной и зависеть от природы алгоритма γ.

Если же множество эталонов Ω немонотонно, то монотонный алгоритм γ, обученный на множестве Ω, не сможет классифицировать все объекты Ω в соответствии с известными метками классов. Кроме того, дефектные пары Ω могут провоцировать ошибки алгоритма γ на других объектах.

Для оценки ошибок нам будут полезны следующие понятия и обозначения:

Опр. 3.1.5. *Объект u **доминирует**, над объектом v если:*

- *объекты u и v принадлежат классу 1, и $v > u$*
- *объекты u и v принадлежат классу 0, и $v < u$*

Опр. 3.1.6. *Объект v будем называть **недоминируемым**, если не существует объекта $u \in X^L$, который бы доминировал над объектом v.*

Опр. 3.1.7. *Будем называть **пессимистичным монотонным классификатором с эталонным множеством** Ω и обозначать γ^p_Ω классификатор, который:*

- *правильно классифицирует объект u класса 1 тогда и только тогда, когда существует объект $a \in \mathbb{A} \cap \Omega$ такой, что $a \leq u$, и не существует объекта $b \in \mathbb{B} \cap \Omega$ такого, что $u \leq b$*
- *правильно классифицирует объект v класса 0 тогда и только тогда, когда существует объект $b \in \mathbb{B} \cap \Omega$ такой, что $v \leq b$, и не существует объекта $a \in \mathbb{A} \cap \Omega$ такого, что $a \leq v$*

Обозначим Ω_M подмножество объектов Ω не участвующих в дефектных парах. Пессимистичный монотонный классификатор будет правильно классифицировать те и только те объекты, которые правильно классифицировал бы

любой монотонный алгоритм, работающий корректно на множестве объектов Ω_M. То есть качество работы пессимистичного монотонного классификатора позволяет получить оценку снизу для качества работы монотонных классификаторов, корректно работающих Ω_M.

Покажем, что пессимистичный монотонный классификатор можно представить в виде классификатора ближайшего соседа со специальной функцией расстояния. Это позволит нам использовать формулы подсчета функционала полного скользящего контроля (CCV), известные для алгоритма ближайшего соседа и оценивать таким образом обобщающую способность монотонных классификаторов.

Данная специальная функция расстояний является упрощенным вариантом монотонной функции расстояний [22–24] и выглядит следующим образом:

$$\rho(x,z) = \begin{cases} 0, & \text{если } x \geq z \text{ и } z \in \mathbb{A} \\ 0, & \text{если } x \leq z \text{ и } z \in \mathbb{B} \\ \infty, & \text{в противном случае.} \end{cases}$$

Используя алгоритм ближайшего соседа с данной функцией расстояний и эталонным множеством Ω, предположив дополнительно, что неоднозначность классификации является ошибкой, получим что данный алгоритм

- правильно классифицирует объект u класса 1 тогда и только тогда, когда существует объект $a \in \mathbb{A} \cap \Omega$ такой, что $a \leq u$, и не существует объекта $b \in \mathbb{B} \cap \Omega$ такого, что $u \leq b$
- правильно классифицирует объект v класса 0 тогда и только тогда, когда существует объект $b \in \mathbb{B} \cap \Omega$ такой, что $v \leq b$, и не существует объекта $a \in \mathbb{A} \cap \Omega$ такого, что $a \leq v$

То есть совпадает с пессимистичным монотонным классификатором с эталонным множеством Ω. Следовательно, оценки обобщающей способности алгоритма ближайшего соседа с предложенной функцией расстояния ρ позволяют оценить обобщающую способность семейства монотонных классификаторов.

3.2 Отбор эталонов и признаков на монотонной выборке

Предположим, выборка X^L является монотонной. В этой ситуации отбор объектов и признаков может быть необходим с точки зрения выполнения гипотезы компактности. Кроме того, большое количество неинформативных признаков способствует переобучению, а также делает почти все объекты обучающей выборки несравнимыми. Отсутствие возможности сравнивать объекты не позволяет использовать явный учет априорных предположений о монотонности для повышения обобщающей способности алгоритмов.

Рассмотрим задачи отбора объектов и признаков с условием сохранения ограничения монотонности.

3.2.1 Задача отбора признаков

Задача 3.2.1.1. *Выбрать признаки так, чтобы при отсутствии дефектных пар получить максимальное возможное при этом количество монотонных пар:* $FS(|D| = 0\colon |M| \to \max)$.

Теорема 3.2.1. *Решение задача о покрытии множества подмножествами сводится к решению задачи 3.2.1.1 на некоторой искусственной обучающей выборке* X^L.

Доказательство. Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots C_{|S|}\}$ — семейство его подмножеств.

Будем считать, что ни одно из множеств C_j не содержит все элементы U. (Иначе задача о покрытии множества подмножествами решалась бы тривиально.)

Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Обозначим данное множество объектов как $\mathbb{A}_U$.

Добавим один объект b класса $\mathbb{B}$ и один объект a класса $\mathbb{A}$.

Каждому элементу $C_j \in S$ поставим в соответствие объект $b_j^* \in \mathbb{B}$ и признак f_j. бозначим данное множество объектов как $\mathbb{B}_S^*$.

Пусть $a_i > b$ по признаку f_j тогда и только тогда, когда $u_i \in C_j$. Для каждого признака f_j обозначим данное множество объектов как $\mathbb{A}_j$.

В соответствии со сказанным выше, пусть признак f_j задает следующий порядок:

$$f_j(a) < \left\{ f_j(b_t) \;\middle|\; b_t \in \mathbb{B}_S^* \setminus \{b_j^*\} \right\} < \\ < \left\{ f_j(a_t) \;\middle|\; a_t \in \mathbb{A}_U \setminus \{\mathbb{A}_j\} \right\} < \\ < F_j(b) < \left\{ f_j(a_t) \;\middle|\; a_t \in \mathbb{A}_j \right\} < f_j(b_j^*)$$

Добавим также признак f^*, задающий порядок

$$\left\{ f^*(b_t) \;\middle|\; b_t \in \mathbb{B}_S^* \right\} < \left\{ f^*(a_t) \;\middle|\; a_t \in \mathbb{A}_U \right\} < f^*(b) < f^*(a)$$

Таким образом объект a не участвует в дефектных парах тогда и только тогда, когда множество признаков F содержит признак f^*.

Дефектные пары с участием объектов a_i отсутствуют при множестве признаков F тогда и только тогда, когда для каждого объекта a_i существует признак $f_j \in F$ такой, что $u_i \in C_j$.

Таким образом дефектные пары отсутствую тогда и только тогда, когда множество признаков F содержит признак f^*, а подмножества C_j, соответствующие признакам $f_j \in F$, образуют покрытие.

По построению количество монотонных пар равно $|U| \cdot (|S| - |F|)$. Таким образом количество монотонных пар максимально тогда и только тогда, когда соответствующее покрытие множества подмножествами минимально.

То есть задача о покрытии множества подмножествами сводится к решению задачи 3.2.1.1 на предложенной обучающей выборке, что и требовалось доказать.

Количество объектов и признаков полиномиально по $|U|$ и $|S|$.

Следствие 3.2.1.1. *Задача* $FS(|D| = 0 \colon |M| \geq m)$ *является NP-трудной.*

Задача 3.2.1.2. *Выбрать минимальное количество признаков таким образом, чтобы дефектные пары отсутствовали:* $FS(|D| = 0 \colon |F| \to \min)$

Теорема 3.2.2. *Решение задача о покрытии множества подмножествами сводится к решению задачи 3.2.1.2 на некоторой искусственной обучающей выборке* X^L.

Доказательство. Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots C_{|S|}\}$ — семейство его подмножеств.

Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Обозначим данное множество объектов как $\mathbb{A}_U$. Добавим один объект b класса $\mathbb{B}$.

Каждому подмножеству $C_j \in S$ поставим в соответствие признак F_j.

Пусть $a_i > b$ по признаку f_j тогда и только тогда, когда $u_i \in C_j$. Для каждого признака f_j обозначим данное множество объектов как $\mathbb{A}_j$. Тогда признак f_j задает следующий порядок:

$$\left\{ f_j(a_t) \;\middle|\; a_t \in \mathbb{A}_U \setminus \{\mathbb{A}_j\} \right\} < f_j(b) < \left\{ f_j(a_t) \;\middle|\; a_t \in \mathbb{A}_j \right\}.$$

Дефектные пары отсутствуют при множестве признаков F тогда и только тогда, когда для каждого объекта a_i существует признак $f_j \in F$ такой, что $u_i \in C_j$.

Таким образом дефектные пары отсутствуют тогда и только тогда, когда множества C_j, соответствующие выбранным признакам, образуют покрытие. Покрытие минимальное тогда и только тогда, когда выбранно минимальное множество признаков, обеспечивающих отсутствие дефектных пар.

То есть задача о покрытии множества подмножествами сводится к решению задачи 3.2.1.2 на предложенной обучающей выборке, что и требовалось доказать.

Количество объектов и признаков полиномиально по $|U|$ и $|S|$.

Следствие 3.2.1.2. *Задача* $FS(|D| = 0 \colon |F| \le q)$ *является NP-трудной.*

3.2.2 Задача отбора объектов

Задача 3.2.2.1. *Выбрать из* X^L *минимальное множество эталонных объектов* Ω*, такое что пессимистичный монотонный классификатор с эталон-*

ным множеством Ω *будет классифицировать все объекты* X^L *правильно, то есть частота ошибок будет равна нулю:* $\nu(\gamma^p_\Omega, X^L) = 0$

Теорема 3.2.3. *Минимальное множество эталонных объектов* $\Omega \subseteq X^L$ *такое, что* $\nu(\gamma^p_\Omega, X^L) = 0$*, является множеством всех недоминируемых объектов* X^L*.*

Доказательство. Предположим, множество эталонов Ω не содержит недоминируемый объект $u \in X^L$. Тогда по определению пессимистичного монотонного классификатора он будет ошибаться на объекте u, а значит $\nu(\gamma^p_\Omega, X^L) \neq 0$. Следовательно, все недоминируемые объекты X^L принадлежат Ω.

Предположим теперь, что существует объект $v \in \Omega$ не являющийся недоминируемым. Тогда существует недоминируемый объект $u \in X^L$, который над ним доминирует. Множество эталонов Ω содержит все недоминируемые объекты, следовательно $u \in \Omega$. Тогда над любым объектом, над которым доминирует v, доминирует и u. Следовательно, если произвольный объект x монотонной выборки X^L правильно классифицируется пессимистичным монотонным классификатором γ^p_Ω, то он будет правильно классифицироваться и пессимистичным монотонным классификатором $\gamma^p_{\Omega\setminus v}$.

Следовательно, минимальное множество эталонных объектов $\Omega \subseteq X^L$ такое, что $\nu(\gamma^p_\Omega, X^L) = 0$ включает в себя все недоминируемые объекты X^L и только их.

Утв. 3.2.2.1. *Задача поиска недоминируемых объектов выборки* X^L *полиномиальна по количеству объектов.*

Доказательство. Примитивный алгоритм, сравнивающий объекты попарно и отбрасывающий доминируемые работает $O(|X^L|^2)$.

Следствие 3.2.2.1. *Задача поиска минимального множества эталонных объектов* $\Omega \subseteq X^L$ *такого, что* $\nu(\gamma^p_\Omega, X^L) = 0$*, является полиномиальной по количеству объектов в выборке.*

В отличии от функционала $\nu(\gamma^p_\Omega, X^L)$ для пессимистичного монотонного классификатора функционал $Q_k(\mu_\Omega)$ всегда будет отличен от нуля за счет вклада недоминируемых объектов.

Задача 3.2.2.2. *Выбрать из X^L минимальное множество эталонных объектов Ω, такое что значение функционала $Q_k(\mu_\Omega)$ пессимистичного монотонного классификатора с эталонным множеством Ω будет минимально возможным.*

Теорема 3.2.4. *Решение задача о покрытии множества подмножествами сводится к решению задачи 3.2.2.2 на некоторой искусственной обучающей выборке X^L.*

Доказательство.

Рассмотрим $Q_1(\mu_\Omega) = C(1)P(1)$ Тогда минимально возможное значение $Q_1(\mu_\Omega)$ равно вкладу в функционал недоминируемых объектов.

Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots C_{|S|}\}$ — семейство его подмножеств.

Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Обозначим данное множество объектов как $\mathbb{A}_U$. Каждому подмножеству $C_j \in S$ поставим в соответствие объект $a_j^* \in \mathbb{A}$. Обозначим данное множество объектов как $\mathbb{A}_S^*$ Добавим один объект b класса $\mathbb{B}$.

Пусть $a_j^* < a_i$ тогда и только тогда, когда $u_i \in C_j$. Для каждого a_i обозначим данное множество объектов как $\mathbb{A}_i^*$.

Все остальные объекты попарно несравнимы.

Покажем, что такую выборку X^L действительно можно построить. Для этого построим множество признаков, задающих соответствующий частичный порядок:

Для каждого объекта a_i зададим признак f_i такой, что он задает следующий порядок:

$$\left\{ f_i(a) \;\middle|\; a \in \mathbb{A}_U \setminus \{a_i\} \right\} > \left\{ f_i(a) \;\middle|\; a \in \mathbb{A}_S^* \setminus \mathbb{A}_i^* \right\} > f_i(a_i) > \left\{ f_i(a) \;\middle|\; a \in \mathbb{A}_i^* \right\}$$

И два вспомогательных признака f_1^*

$$f_1^*(a_1) > f_1^*(a_2) > \ldots > f_1^*(a_{|U|}) > f_1^*(a_1^*) > \ldots > f_1^*(a_{|S|}^*) > f_1^*(b)$$

и f_2^*

$$f_2^*(b) > f_2^*(a_{|U|}) > \ldots > f_2^*(a_1) > f_2^*(a^*_{|S|}) > \ldots > f_2^*(a_1^*)$$

Тогда все объекты a_j^* и b являются недоминируемыми, а значит при любом выборе эталонов пессимистичный монотонный классификатор будет на них ошибаться.

Тогда минимальное значение $Q_1(\mu_\Omega) = \frac{|S|+1}{|U|+|S|+1}$ и достигается тогда и только тогда, когда для любого a_i существует $a_j^* \in \Omega$ такое, что $a_j^* < a_i$, то есть для любого $u_i \in U$ существует $C_j \in S$ такое, что $u_i \in C_j$ и $a_j^* \in \Omega$. Данные условия выполняются только если подмножества C_j, соответствующие $a_j^* \in \Omega$, образуют покрытие множества U. Минимальное множество Ω соответствует минимальному покрытию множества U подмножествами $C_j \in S$.

Таким образом решение задача о покрытии множества подмножествами сводится к решению задачи 3.2.2.2 на некоторой искусственной обучающей выборке X^L, что и требовалось доказать.

Следствие 3.2.2.2. *Задача выбора из X^L множества эталонных объектов Ω, такого что $|\Omega| \leq t$ и значение функционала $Q_1(\mu_\Omega)$ пессимистичного монотонного классификатора с эталонным множеством Ω будет минимально возможным, является NP-трудной.*

Рассмотрим теперь задачи отбора объектов и признаков с точки зрения монотонизации выборки.

3.3 Задача монотонизации выборки

Утв. 3.3.0.1. *Для произвольного подмножества признаков $F \subseteq \mathbb{F}$*

$$M_F = \bigcap_{f \in F} M_f, \quad D_F = \bigcap_{f \in F} D_f.$$

Следствие 3.3.0.1. *Для любых подмножеств признаков F, F'*

$$F \subseteq F' \quad \Rightarrow \quad M_{F'} \subseteq M_F, \ D_{F'} \subseteq D_F \quad \Rightarrow \quad |M_{F'}| \leq |M_F|, \ |D_{F'}| \leq |D_F|. \quad (3.1)$$

Задача построения монотонного классификатора заключается в аппроксимации неизвестной функции $y^*\colon \mathbb{X} \to \{0,1\}$, заданной в точках обучающей выборки, монотонной функцией $y\colon \mathbb{X} \to \{0,1\}$. По определению монотонной функции для любых двух объектов $x,x' \in \mathbb{X}$ из $x < x'$ следует $y(x) \leq y(x')$. На практике обучающая выборка может не удовлетворять условию монотонности, и в таких случаях ставится задача её предварительной монотонизации.

Задача монотонизации обучающей выборки состоит в том, чтобы выбрать подмножество объектов $\Omega \subseteq X^L$ и подмножество признаков $F \subseteq \mathbb{F}$ так, чтобы среди пар объектов (a,b) из $(\mathbb{A} \cap \Omega) \times (\mathbb{B} \cap \Omega)$ оказалось как можно меньше дефектных пар и как можно больше монотонных: $|D_F| \to \min$, $|M_F| \to \max$. Согласно (3.1), первое условие можно заменить косвенным требованием $|F| \to \max$ или $|\Omega| \to \min$, а второе — косвенным требованием $|F| \to \min$ или $|\Omega| \to \max$.

Заметим, что минимизация $|\Omega|$ при сохранении качества классификации соответствует задаче отбора эталонных объектов (prototype selection, PS) [4;38; 50], а максимизация $|\Omega|$ — задаче отсева выбросов (outlier detection) [51]. Требование максимизации числа признаков $|F|$ возникает, когда априори известно, что все признаки информативны, а требование минимизации $|F|$ соответствует задаче отбора признаков (feature selection, FS), когда требуется отбросить шумовые признаки и уменьшить переобучение [4]. Таким образом, дополнительные условия оптимальности могут частично противоречить друг другу. Для смягчения этих противоречий требования максимизации или минимизации могут заменяться ограничениями-неравенствами.

Систематизация возникающих при этом постановок задач показана на рис. 3.1. Стрелками соединены те задачи, которые получаются друг из друга изменением одного из условий оптимальности. Вычислительная сложность задачи отбора объектов (PS) и отбора признаков (FS) рассматриваются отдельно.

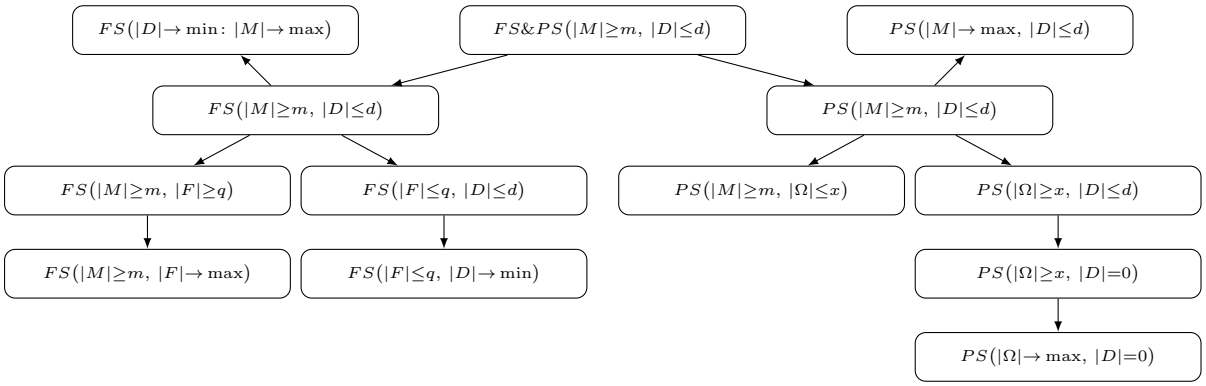


Рисунок 3.1 — Постановки задачи монотонизации: FS — отбор признаков, PS — отбор объектов.

3.3.1 Задача отбора признаков

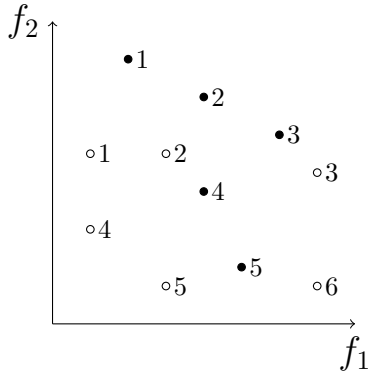
Введем следующее представление множества объектов обучающей выборки X^L: каждому признаку f поставим в соответствие матрицу Ψ^f_{**}, строки которой соответствуют элементам множества $\mathbb{A}$, столбцы — элементам $\mathbb{B}$. Клетки матрицы соответствуют парам объектов из разных классов. В клетке запишем 1, если объект класса $\mathbb{A}$, соответствующий строке, больше объекта класса $\mathbb{B}$, соответствующего столбцу, и 0 в противном случае, то есть $\Psi^f_{ab} = [f(a) > f(b)]$ для всех $(a,b) \in \mathbb{A} \times \mathbb{B}$.

Частичному порядку, индуцированному подмножеством признаков $F \subseteq \mathbb{F}$, соответствует матрица Ψ^F_{**}. В каждой клетке Ψ^F_{**} записывается число признаков из F, по которым объект a, соответствующий строке, больше объекта b, соответствующего столбцу: для всех $(a,b) \in \mathbb{A} \times \mathbb{B}$

$$\Psi^F_{ab} = \sum_{f \in F} [f(a) > f(b)].$$

Замечание 3.3.1. *В общем случае соответствие между частичным порядком и матрицей Ψ^F_{**} на $\mathbb{A}\times\mathbb{B}$ не является взаимнооднозначным: по заданному частичному порядку матрица строится единственным образом, но одна и та же матрица может соответствовать нескольким различным частичным порядкам.*

Пример 3.3.1. *Двумерная выборка и соответствующая ей матрица $\Psi^{\{f_1,f_2\}}_{**}$ показаны на рис. 3.2.*



	b_1	b_2	b_3	b_4	b_5	b_6
a_1	2	1	1	2	1	1
a_2	2	2	1	2	2	1
a_3	2	2	1	2	2	1
a_4	1	1	0	2	2	1
a_5	1	1	0	1	2	1

Рисунок 3.2 — Пример выборки и матрицы: a_i — черные, b_j — белые.

Для дефектных пар в соответствующей клетке матрицы находится 0; для монотонных пар — число признаков $|F|$, то есть

$$\begin{aligned}(a,b) \in D_F &\quad\Leftrightarrow\quad \Psi_{ab}^F = 0;\\ (a,b) \in M_F &\quad\Leftrightarrow\quad \Psi_{ab}^F = |F|.\end{aligned}$$

Задача 3.3.1.1. *Выбрать признаки так, чтобы при минимальном количестве дефектных пар получить максимальное возможное при этом количество монотонных пар:* $FS(|D| \to \min\colon |M| \to \max)$.

Теорема 3.3.1. *Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 3.3.1.1 на некоторой искусственной обучающей выборке* X^L.

Доказательство. Пусть U — конечное множество, $|U| = n$, $S = \{C_1, \dots C_{|S|}\}$ — семейство его подмножеств.

Каждому элементу $u_i \in U$ поставим в соответствие объект $a_i \in \mathbb{A}$. Каждому подмножеству C_t поставим в соответствие объект $b_{t+1} \in \mathbb{B}$ и признак f_{t+1}. Добавим два вспомогательных объекта a^M и a^* класса $\mathbb{A}$, один вспомогательный объект b_1 класса $\mathbb{B}$ и один вспомогательный признак f_1.

Первый признак f_1 задает следующий порядок на множестве объектов:

$$f_1(a^M) > f_1(b_1) > f_1(a_1) > \ldots > f_1(a_n) > f_1(a^*) > f_1(b_2) > \ldots > f_1(b_{|S|+1}).$$

Матрица $\Psi_{**}^{f_1}$ будет выглядеть следующим образом: столбцы упорядочены по номерам элементов класса $\mathbb{B}$, строки — по номерам элементов класса $\mathbb{A}$, две нижние строки соответствуют объектам a^* и a^M. Последняя строка состоит

из единиц, первый столбец, кроме последнего элемента — нули, остальные — единицы, рис. 3.3.

	b_1	b_2			$b_{\lvert S\rvert+1}$
a_1	0	1	1	1	1
a_2	0	1	1	1	1
a_n	0	1	1	1	1
a^*	0	1	1	1	1
a^M	1	1	1	1	1

Рисунок 3.3 — Матрица для первого признака.

Остальные признаки строятся следующим образом: признак с номером $t+1$ соответствует подмножеству $C_t = \{u_{t1}, \ldots, u_{t|C_t|}\}$ из семейства S. Он устанавливает следующий порядок на множестве объектов обучающей выборки:

$$\Big\{ f_{t+1}(a_r) \;\Big|\; a_r \in \{a_{t1}, \ldots, a_{t|C_t|}\} \Big\} > f_{t+1}(b_1) >$$
$$> \Big\{ f_{t+1}(a_r) \;\Big|\; a_r \in \{a_1, \ldots, a_n\} \backslash \{a_{t1}, \ldots, a_{t|C_t|}\} \Big\} > f_{t+1}(b_{t+1}) >$$
$$> f_{t+1}(a^*) > \Big\{ f_{t+1}(b_r) \;\Big|\; b_r \in \mathbb{B} \backslash \{b_1, b_{t+1}\} \Big\} > f_{t+1}(a^M). \quad (3.2)$$

В матрице $\Psi_{**}^{f_{t+1}}$ последняя строка, соответствующая a^M, состоит из нулей; в строке, соответствующей a^*, нули стоят в 1-м и в $(t{+}1)$-м столбце, остальные элементы — единицы; в первом столбце единицы стоят только в строках $\{a_{t1}, \ldots, a_{t|C_t|}\}$, остальные элементы — нули; остальные элементы матрицы — единицы, рис. 3.4.

	b_1	b_2			b_{t+1}			$b_{\lvert S\rvert+1}$
a_1	0	1	1	1	1	1	1	1
a_2	0	1	1	1	1	1	1	1
a_{ti}	1	1	1	1	1	1	1	1
a_n	0	1	1	1	1	1	1	1
a^*	0	1	1	1	1	0	1	1
a^M	0	0	0	0	0	0	0	0

Рисунок 3.4 — Матрица для признака $t+1$.

Итак, построены $|S|+1$ признака. Первый признак — вспомогательный, каждый признак с номером $t+1$ $(t>0)$ взаимно однозначно соответствует подмножеству $C_t \subset S$. Соответствие устанавливается по первому столбцу мат-

рицы: единицы стоят в тех и только в тех строках, которые соответствуют элементам C_t.

Далее для доказательства теоремы нам понадобится несколько вспомогательных утверждений.

Утв. 3.3.1.1. *Оптимальное множество признаков F содержит первый признак f_1.*

Доказательство. Предположим, было выбрано оптимальное подмножество признаков F. Пусть первый признак ему не принадлежит. Рассмотрим матрицу Ψ^F_{**}. Она равна поэлементной сумме матриц всех признаков. При этом нули соответствуют дефектным парам объектов, а элементы со значением $|F|$ — монотонным парам. Поскольку первый признак не принадлежит F, последняя строка Ψ^F_{**} будет состоять из нулей. Добавим к множеству F первый признак $F' = F \cup \{f_1\}$. Рассмотрим матрицу $\Psi^{F'}_{**}$. Матрица $\Psi^{F'}_{**}$ равна сумме матрицы Ψ^F_{**} и матрицы $\Psi^{f_1}_{**}$ первого признака. Таким образом, нулей в матрице $\Psi^{F'}_{**}$ строго меньше, чем в Ψ^F_{**}, то есть дефектных пар по подмножеству признаков F' строго меньше, чем по подмножеству F, что противоречит минимальности количества дефектных пар, а значит и оптимальности F. Таким образом, признак f_1 обязательно принадлежит оптимальному подмножеству признаков. Утверждение 3.3.1.1 доказано. □

Утв. 3.3.1.2. *Матрица Ψ^F_{**} оптимального подмножества признаков F содержит ровно один нулевой элемент $\Psi^F_{a^*b_1} = 0$.*

Доказательство. Матрица любого подмножества признаков является суммой матриц каждого из признаков. Во всех матрицах признаков в первом столбце в строке a^* находится нуль. Следовательно, для любого подмножества признаков F элемент $\Psi^F_{a^*b_1}$ равен нулю. Рассмотрим матрицу всех признаков $\Psi^{\mathbb{F}}_{**}$. Из предположения о существовании покрытия $\bigcup\limits_{C_i \in S} C_i = U$ следует, что для каждого элемента первого столбца $\Psi^{\mathbb{F}}_{ab_1}$, кроме $\Psi^{\mathbb{F}}_{a^*b_1}$, найдется такой признак f_i, что $\Psi^{f_i}_{ab_1} = 1$. По построению, $\Psi^{f_1}_{a^M b_1} = 1$. Таким образом, в матрице $\Psi^{\mathbb{F}}_{**}$ ровно один нулевой элемент, $\Psi^{\mathbb{F}}_{a^*b_1} = 0$. Следовательно, по условию оптимальности множества признаков F, в матрице Ψ^F_{**} ровно один нулевой элемент $\Psi^F_{a^*b_1} = 0$. Утверждение 3.3.1.2 доказано. □

Утв. 3.3.1.3. *Подмножества C_i, соответствующие признакам оптимального множества F, образуют покрытие множества U.*

Доказательство. По построению, каждому элементу $u_i \in U$ соответствует одна из строк матрицы Ψ^F_{**} с соответствующим номером $i = 1, \ldots, n$. Согласно утверждению 3.3.1.2, в матрице Ψ^F_{**} ровно один нулевой элемент, причем он находится в первом столбце в строке a^*. Следовательно, для каждой строки с номером $i = 1, \ldots, n$ существует признак f_{j+1} такой, что $\Psi^{f_{j+1}}_{a_i b_1} = 1$. Тогда в соответствующем множестве C_j содержится элемент u_i. То есть для каждого элемента $u_i \in U$ найдется такое подмножество C_j, соответствующее признаку $f_{j+1} \in F$, что $u_i \in C_j$. Таким образом, подмножества C_i, соответствующие признакам оптимального множества F, образуют покрытие множества U. Утверждение 3.3.1.3 доказано. □

Итак, мы доказали, что оптимальное множество признаков F обязательно содержит первый признак, и что множество подмножеств, соответствующих признакам из F, являются покрытием множества U. Теперь рассмотрим вопрос минимальности данного покрытия. Для этого рассмотрим матрицу $\Psi^{f_{t+1}}_{**}$ признака $t+1$, соответствующего подмножеству C_t. В этой матрице на пересечении строки a^* и столбца b_{t+1} стоит ноль, $\Psi^{f_{t+1}}_{a^* b_{t+1}} = 0$. Рассмотрим матрицу Ψ^F_{**}: первый столбец, кроме элемента $\Psi^F_{a^* 1} = 0$, состоит из элементов, больших 0 и меньших $|F|$, то есть соответствующих парам несравнимых объектов. Элемент $\Psi^F_{a^* 1} = 0$ соответствует единственной дефектной паре. Строка a^M состоит из единиц, значит соответствуют парам несравнимых объектов. В пересечении всех строк с номерами $i = 1, \ldots, n$ и столбцов с номерами $j = 2, \ldots, |S|+1$ стоят элементы, равные $|F|$, то есть соответствующие монотонным парам. Рассмотрим единственную оставшуюся строку a^*: первый элемент этой строки равен нулю; остальные элементы этой строки равны $|F|$ или $|F| - 1$, то есть соответствуют или монотонной или несравнимой паре. По построению, элемент $\Psi^F_{a^* b_{t+1}}$ равен $|F| - 1$ тогда и только тогда, когда признак f_{t+1}, соответствующий подмножеству C_t, принадлежит F, то есть тогда и только тогда, когда $f_{t+1} \in F$. Таким образом, в строке a^* ровно один нулевой элемент, соответствующий дефектной паре; ровно $|F| - 1$ элемент, равный $|F| - 1$ и соответствующий несравнимой паре; остальные элементы равны $|F|$ и соответствуют монотонным парам.

Утв. 3.3.1.4. *Подмножества C_{i-1}, соответствующие признакам $f_i \in F \setminus \{f_1\}$ оптимального подмножества признаков F, образуют* минимальное *покрытие C множества U.*

Доказательство. То, что соответствующие подмножества C_i образуют покрытие C множества U, мы доказали в утверждении 3.3.1.3. Покажем теперь, что покрытие C минимально. Пусть F — оптимальное подмножество признаков. Предположим, что соответствующее покрытие U не является минимальным, то есть существует покрытие C' множества U подмножествами $C_i \in S$, содержащее меньше $|F| - 1$ элементов. Построим множество признаков F' следующим образом: пусть оно состоит из первого признак f_1 и всех признаков, соответствующих подмножествам C_i из минимального покрытия.

Поскольку C' является покрытием, в матрице $\Psi_{**}^{F'}$ первый столбец кроме строки a^* состоит из элементов, больших 0 и меньших $|F'|$, то есть соответствующих парам несравнимых объектов. Элемент, находящийся на пересечении первого столбца и строки a^*, является нулем и соответствует единственной дефектной паре. Строка a^M состоит из единиц, а единицы больше 0 и меньше $|F'|$, значит, соответствуют парам несравнимых объектов. В пересечении всех строк с номерами $i = 1, \ldots, n$ и столбцов с номерами $j = 2, \ldots, |S|+1$ стоят элементы, равные $|F'|$, то есть соответствующие монотонным парам. По построению, в строке a^* ровно один нулевой элемент, соответствующий дефектной паре; ровно $|F'| - 1$ элемент, равный $|F'| - 1$ и соответствующий несравнимой паре; остальные элементы равны $|F'|$ и соответствуют монотонным парам. По предположению, $|F'| < |F|$. Тогда множество признаков F' индуцирует на обучающей выборке порядок, при котором дефектная пара одна, а несравнимых пар меньше, чем в порядке, индуцируемом F. То есть дефектных пар столько же, а монотонных пар больше, что противоречит оптимальности F. Следовательно, C является минимальным покрытием множества U подмножествами. Утверждение 3.3.1.4 доказано. □

Таким образом мы получили конструктивное доказательство теоремы 3.3.1. □

Задача 3.3.1.2. *Выбрать множество признаков F так, чтобы монотонных пар было не менее m, а дефектных не более d: $FS(|D_F| \leq d, |M| \geq m)$.*

Теорема 3.3.2. *Задача $FS(|D_F| \leq d, |M| \geq m)$ является NP-трудной.*

Доказательство. Предположим, что для задачи 3.3.1.2 имеется полиномиальный алгоритм. Построим тогда алгоритм, решающий задачу при всех возможных параметрах d и m. Поскольку d и m могут принимать целые значения от 0 до $|\mathbb{A}||\mathbb{B}|$, потребуется перебрать не более $(|\mathbb{A}||\mathbb{B}|+1)^2$ вариантов. То есть получаемый алгоритм также является полиномиальным, но решает задачу $FS(|D| \to \min\colon |M| \to \max)$. Но тогда полученный алгоритм решал бы за полиномиальное время и задачу о минимальном покрытии множества подмножествами.

Значит, задача 3.3.1.2 является NP-трудной. Теорема доказана. □

Задача 3.3.1.3. *Получить не менее m монотонных пар, выбрав не менее q признаков: $FS(|M| \geq m, |F| \geq q)$.*

Теорема 3.3.3. *Задача $FS(|M| \geq m, |F| \geq q)$ является NP-трудной.*

Доказательство. Сведем задачу о поиске биклики $K_{t,t}$ к данной задаче.

Пусть дан двудольный граф G. Поставим в соответствие каждой вершине левой доли признаки, а каждой вершине правой доли - пары объектов, принадлежащих различным классам. Пусть пары вершин, связанные в G ребром, будут соответствовать признаку и паре объектов, монотонной по данному признаку. Для данного графа обозначим V - множество вершин, L - множество вершин левой доли, R - множество вершин правой доли. Покажем, что такую подвыборку всегда можно построить:

Построим выборку, состоящую из одного объекта a класса 1 и $|R|$ объектов $\{b_1, \ldots, b_{|R|}\}$ класса 0. Каждой вершине j правой доли $j \in R$ поставим во взаимнооднозначное соответствие пару объектов (a, b_j). Других пар объектов, принадлежащих различным классам, нет.

Каждой вершине i левой доли $i \in L$ поставим во взаимнооднозначное соответствие признак $f_i \in F$, $|F| = |L|$.

Рассмотрим теперь порядок, индуцируемый признаками. Пусть каждый признак $f_i \in F$ индуцирует следующий порядок на объектах: $f_i(a) > f_i(b_j)$, если вершина i левой доли $i \in L$ и вершина j правой доли $j \in R$ соединены ребром, и $f_i(a) < f_i(b_j)$ в противном случае. Получим непротиворечивый частичный порядок. Дополним данный частичный порядок до линейного произвольным образом. Таким образом пара объектов (a, b_j) монотонна по $f_i \in F$ тогда и только тогда, когда вершина i левой доли $i \in L$ и вершина j правой доли $j \in R$ соединены ребром.

Получим искомую подвыборку с описанным выше частичным порядком. Она включает в себя не более $|V|+1$ объектов и не более $|V|$ признаков.

В ней каждому признаку взаимно однозначно соответствует вершина левой доли графа G, а каждой паре объектов различных классов взаимно однозначно соответствует вершина правой доли графа G.

Предположим, что существует полиномиальный алгоритм, решающий задачу выбора не менее q признаков так, чтобы монотонных пар было не менее m.

Тогда построим соответствующую графу G обучающую выборку и запустим алгоритм поиска не менее t признаков так, чтобы монотонных пар было не менее t.

Пара является монотонной тогда и только тогда, когда она монотонна по всем выбранным признакам. То есть каждая вершина правой доли, соответствующая монотонной паре, связана со всеми вершинами левой доли, соответствующими выбранным признакам. Следовательно, найденное подмножество признаков и монотонных по ним пар будут соответствовать биклике.

Тогда из предположения существования полиномиального алгоритма для нашей задачи 3.3.1.3 следует существование полиномиального алгоритма для поиска биклики. Таким образом, задача 3.3.1.3 является NP-трудной. Теорема доказана.

□

Задача 3.3.1.4. *Получить минимальное количество дефектных пар, выбрав не более q признаков:* $FS(|D_F| \to \min, |F| \leq q)$.

Теорема 3.3.4. *Решение задачи о минимальном покрытии множества подмножествами сводится к решению задачи 3.3.1.4 на некоторой искусственной обучающей выборке X^L.*

Доказательство. Пусть U — конечное множество, $|U| = n$, S — семейство его подмножеств. Построим обучающую выборку X^L, которая состоит из n объектов класса 1 и n объектов класса 0.

Поставим в соответствие каждому элементу u_j множества U клетку $\Psi^F_{a_j b_j}$ диагонали матрицы Ψ^F_{**}. Для краткости главную диагональ $\Psi^F_{a_j b_j}$ матрицы Ψ^F_{**} будем называть просто *диагональю*. Каждому элементу C_i множества S поставим в соответствие признак f_i, матрица $\Psi^{f_i}_{**}$ которого имеет единицы выше

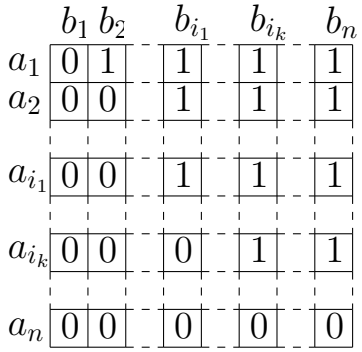


Рисунок 3.5 — Матрица для i-го признака.

диагонали, нули ниже диагонали, а на диагонали единицы в тех и только в тех клетках $\Psi^{f_i}_{a_j b_j}$, которые соответствуют элементам u_j множества C_i.

Далее нам понадобится несколько вспомогательных утверждений.

Утв. 3.3.1.5. *Все элементы матрицы Ψ^F_{**} ниже диагонали равны 0, а выше диагонали равны $|F|$.*

Доказательство. По построению, все элементы матрицы каждого признака выше диагонали равны единице, а ниже диагонали равны нулю. Тогда из равенства $\Psi^F_{ab} = \sum\limits_{f \in F} \Psi^f_{ab}$ следует, что все элементы матрицы Ψ^F_{**} ниже диагонали равны 0, а выше диагонали равны $|F|$. Утверждение 3.3.1.5 доказано. □

Утв. 3.3.1.6. *Все элементы диагонали матрицы Ψ^F_{**} не равны нулю тогда и только тогда, когда множества C_i, соответствующие признакам $f_i \in F$, образуют покрытие множества U, то есть $\bigcup\limits_{i:\, f_i \in F} C_i = U$.*

Доказательство. Поскольку $\Psi^F_{ab} = \sum\limits_{f \in F} \Psi^f_{ab}$, все элементы диагонали не равны нулю тогда и только тогда, когда для каждого $j = 1, \ldots, n$ существует признак $f_i \in F$ такой, что $\Psi^{f_i}_{a_j b_j} = 1$. В то же время $\Psi^{f_i}_{a_j b_j} = 1$, тогда и только тогда, когда $u_j \in C_i$. Следовательно, все элементы диагонали не равны нулю тогда и только тогда, когда для каждого $j = 1, \ldots, n$ существует $f_i \in F$ такой, что $u_j \in C_i$. Тогда $\bigcup\limits_{\{i | f_i \in F\}} C_i = U$. Утверждение 3.3.1.6 доказано. □

Следствие 3.3.1.1. *Пусть F — множество, состоящее не более чем из q признаков, такое, что все элементы диагонали Ψ^F_{**} не равны нулю. Тогда множество подмножеств C_i, соответствующих признакам $f_i \in F$, является решением задачи о покрытии множества подмножествами.*

Осталось показать, что построенные нами матрицы действительно могут соответствовать признакам. Пусть элементы, соответствующие столбцам и строкам, упорядочены: $a_1 > a_2 > \ldots > a_n,\ b_1 > b_2 > \ldots > b_n$. Пусть в соответствующей матрице единицы на диагонали стоят в клетках $i_1, \ldots i_k$. Тогда $a_j > b_j$ для всех $j \in \{i_1, \ldots, i_k\}$ и $a_j < b_j$ для всех $j \notin \{i_1, \ldots, i_k\}$. Причем $a_j > a_{j+1},\ b_j > b_{j+1},\ a_j > b_{j+1},\ b_j > a_{j+1}$ для всех $j = 1, \ldots, n-1$. Таким образом, матрица действительно задает отношение порядка на элементах и является матрицей признака.

Итак, мы свели задачу о минимальном покрытии множества подмножествами к задаче $FS(|D_F| \to \min, |F| \leq q)$.

Теорема 3.3.4 доказана. □

Задача 3.3.1.5. *Получить не более d дефектных пар, выбрав не более q признаков: $FS(|F| \leq q, |D_F| \leq d)$.*

Теорема 3.3.5. *Задача $FS(|F| \leq q, |D_F| \leq d)$ является NP-трудной.*

Доказательство. Предположим, что для задачи 3.3.1.5 имеется полиномиальный алгоритм решения. Число d может принимать целые значения от 0 до $|\mathbb{A}||\mathbb{B}|$. Тогда построим новый алгоритм, решающий задачу 3.3.1.5 для каждого возможного значения d. Он будет решать задачу $FS(|D_F| \to \min, |F| \leq q)$ за полиномиальное время.

Таким образом, задача отбора признаков 3.3.1.5 является NP-трудной. Теорема доказана. □

3.3.2 Задача отбора объектов

Задача 3.3.2.1. *Выбрать объекты так, чтобы монотонных пар было не менее m, а дефектных пар не более d: $PS(|D| \leq d, |M| \geq m)$.*

Гипотеза Задача $PS(|D| \leq d, |M| \geq m)$ является NP-трудной.

К сожалению, доказать данную гипотезу пока не удалось.

Задача 3.3.2.2. *Выбрать не более x объектов так, чтобы число монотонных пар было не менее m: $PS(|M| \geq m, |\Omega| \leq x)$.*

Теорема 3.3.6. *Задача $PS(|M| \geq m, |\Omega| \leq x)$ является NP-трудной.*

Доказательство. Сведем задачу о поиске о поиске биклики $K_{t,t}$ к данной задаче.

Пусть дан двудольный граф G. Поставим в соответствие каждой вершине левой доли объект класса 0, а каждой вершине правой доли — объект класса 1. Пусть пары, связанные в G ребром, будут соответствовать монотонным парам.

Покажем, что такую подвыборку всегда можно построить, то есть построить признаки, индуцирующие соответствующий частичный порядок. Пронумеруем объекты класса 0 и класса 1. Построим два вспомогательных признака f' и f'', индуцирующих два линейных порядка, соответственно:

$$f'(a_1) > f'(a_2) > \ldots > f'(a_{|\mathbb{A}|}) > f'(b_1) > f'(b_2) > \ldots > f'(b_{|\mathbb{B}|});$$
$$f''(b_1) < f''(b_2) < \ldots < f''(b_{|\mathbb{B}|}) < f''(a_1) < f''(a_2) < \ldots < f''(a_{|\mathbb{A}|}).$$

Тогда в частичном порядке, индуцированном этими двумя признаками, все объекты класса 0 будут меньше объектов класса 1, а объекты внутри классов будут несравнимы.

Для каждого объекта a_i класса 1 обозначим $\mathbb{B}_i$ множество объектов класса 0, таких, что соответствующие вершины графа G связаны ребром.

Каждому объекту a_i поставим во взаимнооднозначное соответствие признак f_i, задающий следующий порядок, дополняемый до линейного произвольным образом:

$$\left\{ f_i(a) \;\middle|\; a \in \mathbb{A}\backslash\{a_i\} \right\} > \left\{ f_i(b) \;\middle|\; b \in \mathbb{B}\backslash\mathbb{B}_i \right\} > f_i(a_i) > \left\{ f_i(b) \;\middle|\; b \in \mathbb{B}_i \right\}.$$

Получим искомую подвыборку с описанным выше частичным порядком. Она включает в себя $|V|$ объектов и не более $|V| + 2$ признаков. В ней каждой вершине графа взаимнооднозначно соответствует объект, а каждому ребру — монотонная пара.

Предположим, что существует полиномиальный алгоритм, решающий задачу выбора не более x объектов так, чтобы монотонных пар было не менее m.

Тогда для каждого $t = 1, \ldots, |V|$ построим соответствующую обучающую выборку и запустим алгоритм поиска не более $2t$ объектов так, чтобы моно-

тонных пар было не менее t^2. Найденные подвыборки будут соответствовать бикликам $K_{t,t}$.

Тогда из предположения существования полиномиального алгоритма для нашей задачи следует существование полиномиального алгоритма для поиска биклики. Таким образом, задача 3.3.2.2 является NP-трудной.

Теорема доказана. □

Задача 3.3.2.3. *Устранить все дефектные пары, оставив в обучающей выборке как можно больше объектов:* $PS(|D| = 0, |\Omega| \to \max)$.

Эту задачу можно переформулировать так: устранить все дефектные пары, удалив из обучающей выборки как можно меньше объектов. Построим полиномиальный алгоритм решения задачи 3.3.2.3 и оценим его сложность.

Рассмотрим объекты, участвующие в дефектных парах. Представим их в виде двудольного графа: в одной доле объекты класса $\mathbb{A}$, в другой доле объекты класса $\mathbb{B}$, ребрами соединены объекты, образующие дефектные пары. Тогда задача об удалении минимального количества объектов так, чтобы дефектных пар не было, линейно сводится к задаче о поиске минимального вершинного покрытия для данного двудольного графа.

Пусть d — количество дефектных пар, a_d и b_d — число объектов класса $\mathbb{A}$ и $\mathbb{B}$ соответственно, задействованных в дефектных парах. Тогда время поиска удаляемых объектов составляет $O(d(a_d + b_d)^{0,5})$ по теореме Кёнига.

Задача 3.3.2.4. *Устранить все дефектные пары, выбрав не менее x объектов обучающей выборки:* $PS(|D| = 0, |\Omega| \geq x)$.

Рассмотрим множество объектов $\bar{\Omega}$, являющееся решением задачи 3.3.2.3 на обучающей выборке X^L. Тогда при $x \leq |\bar{\Omega}|$ множество $\bar{\Omega}$ является также и решением задачи 3.3.2.4. Из максимальности $|\bar{\Omega}|$ следует, что при $x > |\bar{\Omega}|$ задача 3.3.2.4 не имеет решения.

Задача 3.3.2.5. *Получить не более d дефектных пар, выбрав не менее x объектов обучающей выборки X^L:* $PS(|D| \leq d, |\Omega| \geq x)$.

Теорема 3.3.7. *Задача 3.3.2.5 является полиномиальной по числу объектов.*

Доказательство. Рассмотрим объекты, участвующие в дефектных парах. Представим их в виде двудольного графа: в одной доле объекты класса $\mathbb{A}$,

в другой объекты класса $\mathbb{B}$, ребрами соединены объекты, образующие дефектные пары. Удалим d ребер всеми возможными способами и сведем задачу к предыдущей. Это возможно сделать $C^d_{|D|}$ способами, то есть за время $O(|\Omega|^{2d})$, полиномиальное по числу объектов в обучающей выборке, но экспоненциальное по d. Теорема доказана. □

3.4 Основные выводы главы 3

1. Оценки обобщающей способности алгоритма ближайшего соседа с предложенной функцией расстояния позволяют оценить обобщающую способность семейства монотонных классификаторов.
2. Все предложенные постановки задачи отбора признаков на монотонной выборке являются NP-трудными.
3. Для задачи отбора эталонных объектов на монотонной выборке с частотой ошибок в качестве целевой функции существует точный полиномиальный алгоритм.
4. Задача отбора эталонных объектов на монотонной выборке с целевой функцией $Q_1(\mu_\Omega)$ является NP-трудной.
5. Все рассматриваемые постановки задачи монотонизации кроме $PS(|D| = 0, |\Omega| \geq x)$ и $PS(|D| \leq d, |\Omega| \geq x)$ являются NP-трудными.
6. Для $PS(|D| = 0, |\Omega| \geq x)$ указан точный эффективный алгоритм решения.
7. Для $PS(|D| \leq d, |\Omega| \geq x)$ существует алгоритм решения, полиномиальный по числу объектов, но экспоненциальный по d.
8. Постановки $PS(|D| = 0, |\Omega| \geq x)$ и $PS(|D| \leq d, |\Omega| \geq x)$ рассматривают задачу монотонизации только с точки зрения отбора объектов.

Из перечисленных выше утверждений следует, что задачу отбора признаков и объектов для построения классификатора с ограничениями монотонности не всегда возможно и целесообразно решать точно, что обосновывает использование приближенных алгоритмов.

Глава 4. Алгоритм монотонизации с одновременным отбором объектов и признаков

Если рассматривать задачи отбора объектов и признаков последовательно, возникает следующая проблема: неинформативные признаки мешают отбирать объекты, делая почти все объекты обучающей выборки несравнимыми, а шумовые объекты могут не позволить удалить неинформативные признаки без нарушения условия монотонности.

Одновременный отбор объектов и признаков предотвращает попадание в часть локальных минимумов, связанных с шумовыми объектами и неинформативными признаками.

В некоторых практических задачах с ограничениями монотонности метка класса может быть монотонно возрастающей по одним признакам и монотонно убывающей по другим. Функция, монотонно убывающая по некоторому признаку f, является монотонно возрастающей по $-f$. Для удобстава будем считать, что метка класса всегда монотонно возрастает по значению признака, просто некоторые признаки требуется предварительно умножить на -1.

Опр. 4.0.1. *Будем говорить, что признак f присутствует в подмножестве $F \subset \mathbb{F}$* ***с инверсией****, если значение данного признака для каждого из объектов умножено на -1. То есть линейный порядок, индуцируемый данным признаком f на множестве объектов, заменен на обратный.* ***Инвертированный*** *признак будем обозначать $-f$.*

При построении алгоритма монотонизации будем считать, что некоторые из признаков могут присутствовать в подмножестве $F \subset \mathbb{F}$ с инверсией. Может оказаться, что значения признаков некоторых объектов совпадают.

В главе 3 использовалась матрица Ψ^F_{**}, описывающая отношение порядка между объектами классов $\mathbb{A}$ и $\mathbb{B}$ обучающей выборки X^L по признакам F. По аналогии с матрицей Ψ^F_{**} построим трехмерное предствление $\overline{\Psi}$, которое будет описывать отношение порядка между объектами классов $\mathbb{A}$ и $\mathbb{B}$ по всем признакам $\mathbb{F}$.

Каждому признаку $f \in \mathbb{F}$ поставим в соответствие матрицу $\overline{\Psi}^f_{**}$, строки которой соответствуют элементам множества $\mathbb{A}$, столбцы — элементам $\mathbb{B}$.

Клетки матрицы соответствуют парам объектов из разных классов. Заполним их следующим образом:

$$\overline{\Psi}^f_{ab} = \begin{cases} 1, & \text{если } f(a) > f(b); \\ 0, & \text{если } f(a) = f(b); \\ -1, & \text{если } f(a) < f(b). \end{cases}$$

для всех $(a,b) \in \mathbb{A} \times \mathbb{B}$.

Получим, что:

- каждому объекту $a \in \mathbb{A}$ соответствует подматрица $\overline{\Psi}^*_{a*}$ представления $\overline{\Psi}$, состоящая из $|\mathbb{B}| \times |\mathbb{F}|$ элементов;
- каждому объекту $b \in \mathbb{B}$ соответствует подматрица $\overline{\Psi}^*_{*b}$ представления $\overline{\Psi}$, состоящая из $|\mathbb{A}| \times |\mathbb{F}|$ элементов;
- каждому признаку $f \in \mathbb{F}$ соответствует подматрица $\overline{\Psi}^f_{**}$ представления $\overline{\Psi}$, состоящий из $|\mathbb{A}| \times |\mathbb{B}|$ элементов;
- если признак f рассматривается с инверсией, соответствующая подматрица $\overline{\Psi}^f_{**}$ представления $\overline{\Psi}$ умножается на -1.

4.1 Общая схема жадного алгоритма монотонизации

Обозначим алгоритм монотонизации $mon(X^L \times \mathbb{F})$. На **вход** алгоритм принимает выборку $X^L \times \mathbb{F}$. **Результатом** работы алгоритма является подвыборка $\Omega \times F$: $\Omega \subseteq X^L$, $F \subseteq \mathbb{F}$. Тогда $\Omega \times F = mon(X^L \times \mathbb{F})$.

Пусть дан функционал W, характеризующий монотонность выборки.

На каждой итерации:

- Оценивается W
- Для каждого объекта $x \in X^L$ оцениваются изменения W при:
  - исключении объекта x из подмножества $\Omega \subseteq X^L$, если $x \in \Omega$
  - включении объекта x в подмножество $\Omega \subseteq X^L$, если $x \notin \Omega$
- Для каждого признака $f \in \mathbb{F}$ оцениваются значения W если:
  - f отсутствует в подмножестве $F \subseteq \mathbb{F}$ с инверсией и без;
  - f присутствует в подмножестве $F \subseteq \mathbb{F}$ без инверсии;
  - f присутствует в подмножестве $F \subseteq \mathbb{F}$ с инверсией;

Рассматриваются соответствующие изменения W.

– вносится одно или несколько наилучших изменений

Итерации выполняются до тех пор, пока W не перестанет улучшаться. После этого возможно применение дополнительных эвристик. Например, случайное изменение состояния одного или нескольких признаков с целью исследования окрестности полученного алгоритмом экстремума.

4.2 Функционалы монотонности

Для работы жадного алгоритма необходимо задать функционал W, характеризующий монотонность выборки. Рассмотрим различные функционалы монотонности и особенности их жадной оптимизации.

Примером функционала, характеризующего монотонность выборки, является, например, **степень монотонности** (degree of monotonicity) [52]

$$DgrMon = \frac{|M|}{|M| + |D|}.$$

Недостатком $DgrMon$ при использовании в жадном алгоритме монотонизации является то, что он может принять свое оптимальное значение $DgrMon = 1$ при наличии большого количества шумовых признаков, делающих почти все пары несравнимыми. Например, если в всего одна пара монотонна и нет дефектных, $DgrMon = 1$. При отсутствии дефектных пар $DgrMon$ никак не отражает наличие и количество сравнимых пар. Количество сравнимых пар отражается **профилем монотонности** [23], который определяются следующим образом: Клином объекта x_i называют множество объектов

$$\Xi_i = \begin{cases} \{x \in X^L : x_i < x, y(x) = 0\}, & \text{если } y(x_i) = 0; \\ \{x \in X^L : x < x_i, y(x) = 1\}, & \text{если } y(x_i) = 1. \end{cases}$$

Профилем монотонности выборки X^L называется функция, равная доле объектов с клином мощности m:

$$M(m) = \frac{1}{L}\sum_{i=1}^{L}[|\Xi_i| = m].$$

Мощность клина $|\Xi_i|$ характеризует глубину погружения объекта x_i в свой класс, то есть зависит от наличия и количества объектов того же класса, сравнимых с данным. Но мощность клина $|\Xi_i|$ объекта x_i не зависит от объектов, сравнимых с x_i, принадлежащих другому классу.

Следовательно, профиль монотонности не зависит от пар объектов, принадлежащих разным классам. В частности, профиль монотонности не зависит от наличия и количества дефектных и монотонных пар. Например, функция $M(m)$ будет совпадать на выборке X^L для любого строго линейного порядка, независимо от того, выполняются ли для него ограничения монотонности.

Степенью немонотонности выборки X^L называется наименьшая частота ошибок, допускаемая на ней монотонными алгоритмами:

$$\zeta(X^L) = \min_{\gamma \in Mon} \nu(\gamma, X^L),$$

где Mon — множество всех монотонных алгоритмов.

Монотонный алгоритм, классифицирующий выборку, решает задачу ее монотонизации: подмножество объектов, на которых алгоритм выдает правильный ответ, является монотонным. Тогда минимальная частота ошибок достигается тогда, когда алгоритм выдает правильный ответ на максимальном возможном подмножестве $\Omega \subset X^L$ объектов. То есть получаемое таким образом подмножество объектов Ω является решением уже рассмотренной задачи $PS(|\Omega| \to \max, |D| = 0)$. Таким образом подсчет степени монотонности выборки возможен за полиномиальное время.

Степень немонотонности выборки $\zeta(X^L)$, обладает тем же недостатком, что и степень монотонности $DgrMon$: принимает свое оптимальное значение $\zeta(X^L) = 0$ при большом количестве шумовых признаков, делающих все пары несравнимыми.

В работе [23] доказана следующая теорема:

Теорема 4.2.1. *Если метод μ минимизирует эмпирический риск в классе всех монотонных функций и степень немонотонности выборки X^L равна ζ, то*

$$Q_k(\mu) \leq Es(X^L, F) = \sum_{m=0}^{\zeta L+k-1} M(m) \sum_{s=\max\{0,m-k+1\}}^{\min\{\zeta L,\ell,m\}} \frac{C_m^s C_{L-1-m}^{\ell-s}}{C_{L-1}^{\ell}},$$

где k — количество контрольных объектов, а $\ell = L - k$.

Оценка Es [18] зависит и от степени немонотонности ζ и от профиля монотонности $M(m)$, то есть и от количества дефектных пар, и от количества сравнимых пар объектов, принадлежащих одному классу. Однако $Es(X^L, F)$ не зависит от количества монотонных пар.

Пример 4.2.1. *Пусть множество признаков F_1 индуцирует на множестве объектов X_1 такой частичный порядок, что дефектные пары отсутствуют, а монотонные присутствуют. Предположим, существует другой набор признаков F_2, такой, что частичный порядок объектов X_1 внутри классов такой же, как и при F_1, но объекты принадлежащие разным классам несравнимы. Тогда $Es(X_1, F_1) = Es(X_2, F_2)$.*

Итак, все рассмотренные выше функционалы, характеризующие монотонность выборки, обладают определенными недостатками с точки зрения их использования в качестве целевой функции жадного алгоритма. Предложим функционал монотонности W, не обладающий перечисленными недостатками. Для его построения привлечем следующие дополнительные соображения:

1. Функционал монотонности W должен монотонно возрастать по количеству монотонных пар объектов.
2. Функционал монотонности W должен монотонно убывать по количеству дефектных пар объектов.
3. Необходимо, чтобы изменение функционала W при изменении подмножеств F и Ω принимало достаточно большое количество различных значений, чтобы осуществлять жадный выбор.
4. Удобно, чтобы W имел параметры, учитывающие несбалансированность классов и неравноценность ошибок первого и второго рода.
5. Желательно, чтобы алгоритм удалял не слишком большую долю объектов выборки X^L.

6. Множество признаков F можно сокращать настолько сильно, насколько это выгодно с точки зрения монотонности. Допустимо даже использование F, состоящего из одного признака.

Всем перечисленным соображениям удовлетворяет:

$$W = \Big(w_a|A| + w_b|B|\Big) + \sum_{a\in A}\sum_{b\in B}\frac{\sum_{f\in F}\overline{\Psi}^f_{ab}}{\sum_{f\in F}|\overline{\Psi}^f_{ab}|} + W_M|M| - W_D|D|,$$

где W_M — поощрение монотонных пар; W_D — штраф на дефектные пары; w_a — вес класса $\mathbb{A}$; w_b — вес класса $\mathbb{B}$; $A = \Omega \cap \mathbb{A}$, $B = \Omega \cap \mathbb{B}$. Данный функционал используется в эксперименте.

4.3 Описание данных, на которых проводится вычислительный эксперимент

Эксперимент проводится на данных задачи диагностики заболеваний по электрокардиограмме (ЭКГ). В качестве объектов выступают 13066 обследований, а в качестве классов — 13 заболеваний, класс «вегетососудистая дистония» и класс «абсолютно здоров», то есть 15 классов. Ниже приведены названия классов и их сокращенные обозначения.

- абсолютно здоров — АЗ
- вегетососудистая дистония — ВД
- гипертоническая болезнь — ГБ
- желчекаменная болезнь — ЖК
- ишемическая болезнь сердца — ИБ
- мочекаменная болезнь — МК
- миома матки — ММ
- сахарный диабет — СД
- узловой зоб щитовидной железы — УЩ
- хронический гастрит гипоацидный — ХГ
- холицистит хронический — ХХ
- анемия железодифицитная — ЭА

– аденома простаты — АП
– аднексит хронический — АХ
– язвенная болезнь — ЯБ

Каждое обследование состоит из одной или нескольких записей ЭКГ. Постановка задачи и технология построения признакового пространства для анализа ЭКГ-сигналов были разработаны Успенским В. М. [53].

Технология Успенского В. М. основана на кодировании изменения параметров R-пиков ЭКГ в символьную последовательность (кодограмму). Рассматриваются такие параметры, как амплитуды пиков R, интервалы между ними T и соответствующие углы $\alpha = \operatorname{arctg} \frac{R}{T}$.

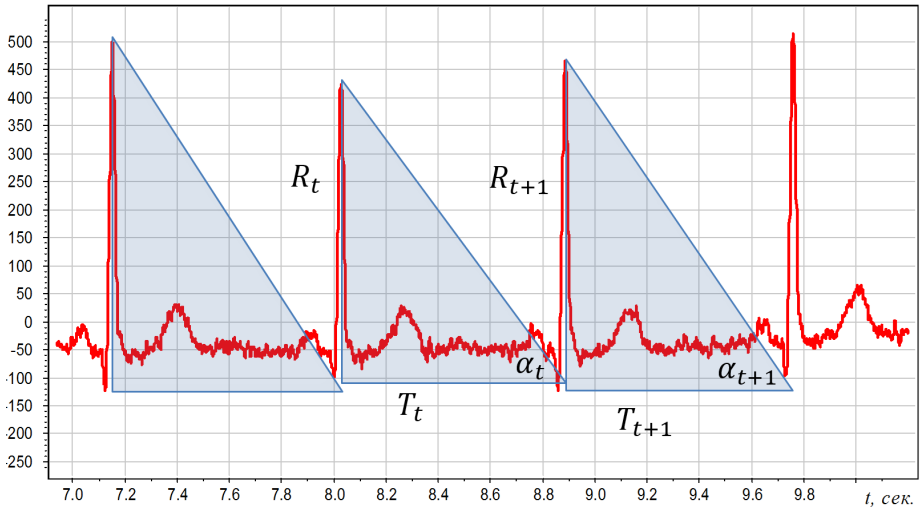


Рисунок 4.1 — параметры R-пиков

Приращения амплитуд, интервалов и углов кодируются по правилу, указанному в таблице 4.1. Пример получаемой кодограммы представлен на рисунке 4.2. В качестве признаков используются частоты вхождений в кодограмму так называемых триграмм — комбинаций из трех последовательных буков. Размерность получаемого признакового пространства $6^3 = 216$.

Таблица 4.1

Правила кодирования

$dR_t = R_{t+1} - R_t$	+	−	+	−	+	−
$dT_t = T_{t+1} - T_t$	+	−	−	+	+	−
$d\alpha_t = \alpha_{t+1} - \alpha_t$	+	+	+	−	−	−
s_t	A	B	C	D	E	F

Рисунок 4.2 — пример кодограммы

Рисунок 4.3 — пример триграммы

4.4 Постановка эксперимента

Конечной целью монотонизации выборки является обучения монотонного алгоритма классификации. Будем оценивать качество алгоритма монотонизации mon при помощи оценки качества соответствующего классификатора γ, обученного на монотонизированной выборке.

В эксперименте для каждой пары классов решается задача двхклассовой классификации в стандартной постановке. Для этого выбираются все записи X^L принадлежащие рассматриваемым классам. Затем множество записей X^L разбивается на обучающее X^ℓ и контрольное X^k подмножества в соотношении примерно $4:1$ с сохранением количественного отношения между объектами различных классов.

На обучающей подвыборке $X^\ell \times \mathbb{F}$ запускается алгоритм монотонизации mon. А на результате его работы $X \times F = mon(X^\ell \times \mathbb{F})$ обучаются алгоритмы классификации, называемые логистическая регрессия с регуляризацией $l1$ и $l2$ [42]. Свойства регуляризации $l1$ обеспечивают дополнительный отбор признаков на этапе классификации. Для сравнения качества классификации аналогичные алгоритмы обучаются на $X^\ell \times \mathbb{F}$ без предварительной монотонизации.

Для измерения качества классификации в медицинской диагностике принято использовать меры чувствительности и специфичности. Обозначим долю ложных положительных классификаций — FPR (False Positive Rate), долю

верных положительных классификаций — TPR (True Positive Rate), долю ложных отрицательных классификаций — FNR (False Negative Rate), долю верных отрицательных классификаций — TNR (True Negative Rate).

Опр. 4.4.1. *Чувствительность – это доля больных, для которых диагностическое правило верно диагностирует наличие болезни.*

$$Чувствительность = \frac{TPR}{TPR + FNR}$$

Опр. 4.4.2. *Специфичность – это доля здоровых, для которых диагностическое правило верно диагностирует отсутствие болезни.*

$$Специфичность = \frac{TNR}{TNR + FPR}$$

Выбор компромисса между чувствительностью и специфичностью зависит от каждого конкретно взятого случая. Потому в данной работе качество классификации предлагается оценивать при помощи ROC-AUC [54].

Опр. 4.4.3. *ROC-AUC (Area Under Curve) — это площадь под кривой, отображающей зависимость чувствительности от специфичности.*

Функционал AUC можно интерпретировать как долю правильно упорядоченных пар прецедентов.

Для каждой задачи классификации, рассмотренной в данном эксперименте, качество классификации оценивается значением AUC, вычисляемом на соответствующей выборке $X^k \times F$. С целью уменьшения влияния разбиения на оценку качества, для каждой задачи классификации эксперимент производится на 5 различных разбиениях, полученные значения AUC усредняются.

4.5 Результаты эксперимента

На серии рисунков 4.4–4.18 изображены признаковые пространства, полученные в результате отбора признаков при помощи алгоритма *mon* на одном из разбиений выборки. Столбцы соответствуют 216 признакам, строки — задаче двуклассовой классификации. Серый цвет означает, что признак был

отброшен алгоритмом, черный цвет — признак присутствует в F без инверсии, белым цветом обозначены признаки, присутствующие в F с инверсией.

Рисунок 4.6 признаковых пространств для класса АЗ (абсолютно здоров) выделяется среди остальных следующим: практически для всех задач двухклассовой классификации с участием класса АЗ отбиралось небольшое количество одних и тех же признаков, присутствующих с одним и тем же знаком. Это означает, что существуют признаки, характерные для АЗ, которых достаточно для отделения класса АЗ от остальных классов. Подобным свойством, выраженным в меньшей мере, обладают рисунки 4.18, 4.8, 4.7 признаковых пространств для класса ВД (вегетососудистая дистония), АХ (аднексит хронический) и ЭА (анемия железодифицитная).

В таблице 4.3 приведены средние значения ROC-AUC на контрольных выборках для следующих алгоритмов: логистическая регрессия с $l1$ и $l2$ регуляризацией, обученная на выборке с предварительной монотонизацией mon и без нее. Доля объектов, отбрасываемых из обучающей подвыборки при монотонизации, составила всреднем около 0,05. С точки зрения прикладной задачи интерес представляют классификаторы, для которых значение ROC-AUC около 0,7 и выше. Строки таблицы 4.3, соответствующие задачам, в которых ROC-AUC не ниже 0.68, отмечены серым. Таких задач 40 из 105, причем в 39 из 40 задач присутствовал хотябы один из классов АЗ, ВД, АХ и ЭА.

Жирным шрифтом выделены значения AUC для тех алгоритмов, для которых предварительная монотонизация обучающей выборки позволила повысить качество классификации. Улучшение качества произошло для 46 из 105 задач. Из 40 задач, в которых было достигнуто значение ROC-AUC не ниже 0.68, улучшение качества произошло в 29 задачах.

В таблице 4.2 для каждого из 15 классов приведено количество задач, в которых участвовал данный класс, и в которых было достигнуто значение ROC-AUC не ниже 0.68. Максимальное возможное количество 14, поскольку в каждой задаче участвовало два различных класса.

Таблица 4.2

Количество задач, в которых участвовал данный класс, и в которых было достигнуто значение ROC-AUC не ниже 0.68

Класс	АЗ	ЭА	АХ	АП	ГБ	ИБ	ХГ	ХХ	МК	ММ	СД	УЩ	ВД	ЯБ	ЖК
Кол-во	**14**	**7**	**9**	4	4	5	4	3	3	3	4	2	**12**	2	4

В приложении А для каждой рассматриваемой задачи двухклассовой классификации приведены графики ROC-кривых соответствующих алгоритмов классификации.

Графики приведены только для одного из 5 разбиений выборки на обучающую и контрольную подвыборки, поскольку они качественно похожи для всех рассматриваемых разбиений. Однако значения ROC-AUC, указанные в легенде, могут незначительно отличаться от средних значений, указанных а таблице 4.3.

Более подробно ход работы алгоритма *mon* отражен на графиках 4.4 и 4.5 (на примере задачи ВД–АЗ). График 4.4 представляет собой зависимость количества монотонных и дефектных пар, количества признаков а также количества объектов каждого из классов в обучающей подвыборке от номера итерации. На графике 4.5 изображены результаты следующего эксперимента: логистическая регрессия с регуляризатором l1 обучалась на обучающей подвыборке, получаемой на каждой итерации алгоритма *mon* (всего 249 классификаторов); для каждого полученного классификатора вычислялось значение ROC-AUC на контрольной подвыборке. Для сравнения на графике приведено значение ROC-AUC на контрольной подвыборке для логистическая регрессия с регуляризатором l1, обученной на обучающей подвыборке без предварительной монотонизации.

Перечислим возможные причины того, что увеличение AUC при использовании алгоритма предварительной монотонизацией *mon* происходит не во всех случаях.

- Отсутствие возможности решить задачу при помощи данного семейства классификаторов.
- Вероятностная природа данных в некоторых задачах может быть такой, что объекты, участвующие в дефектных парах, очень важны при построении разделяющей поверхности между классами. Возможно, некоторые из таких объектов отбрасываются при предварительной монотонизации.
- Переобучение.

Таблица 4.3

Среднее значение AUC на контрольной подвыборке

Классы		|F|	mon_l1	mon_l2	l1	l2
ВД	АЗ	9	**0.862**	**0.863**	0.817	0.799
ГБ	АЗ	9	**0.927**	**0.927**	0.912	0.903
ГБ	ВД	12	**0.721**	**0.721**	0.703	0.701
ЖК	АЗ	6	**0.935**	**0.934**	0.919	0.899
ЖК	ВД	14	**0.701**	**0.701**	0.687	0.685
ЖК	ГБ	31	0.586	0.588	0.607	0.599
ИБ	АЗ	10	0.94	0.94	0.953	0.94
ИБ	ВД	25	0.765	0.766	0.79	0.785
ИБ	ГБ	23	0.546	0.546	0.625	0.626
ИБ	ЖК	33	**0.594**	**0.595**	0.592	0.587
МК	АЗ	9	**0.92**	**0.919**	0.907	0.89
МК	ВД	29	**0.703**	**0.703**	0.683	0.683
МК	ГБ	28	**0.634**	**0.634**	0.627	0.627
МК	ЖК	27	0.561	0.562	0.626	0.627
МК	ИБ	25	0.65	0.65	0.663	0.659
ММ	АЗ	11	**0.906**	**0.905**	0.88	0.873
ММ	ВД	20	0.683	0.683	0.706	0.706
ММ	ГБ	26	0.613	0.612	0.668	0.672
ММ	ЖК	29	0.562	0.561	0.608	0.603
ММ	ИБ	26	0.671	0.671	0.713	0.709
ММ	МК	22	0.548	0.549	0.633	0.634
СД	АЗ	9	**0.933**	**0.932**	0.904	0.894
СД	ВД	23	**0.719**	**0.72**	0.714	0.704
СД	ГБ	32	0.575	**0.575**	0.582	0.573
СД	ЖК	19	**0.487**	**0.486**	0.471	0.474
СД	ИБ	29	0.584	**0.584**	0.588	0.583
СД	МК	26	0.59	0.59	0.636	0.623
СД	ММ	20	0.593	0.593	0.634	0.635
УЩ	АЗ	11	0.91	**0.909**	0.914	0.904
УЩ	ВД	23	**0.705**	**0.703**	0.693	0.682
УЩ	ГБ	37	0.593	0.593	0.673	0.67
УЩ	ЖК	6	0.493	0.492	0.541	0.534
УЩ	ИБ	30	0.613	0.613	0.651	0.647
УЩ	МК	17	0.527	0.527	0.574	0.571
УЩ	ММ	21	0.513	0.512	0.569	0.566
УЩ	СД	34	0.512	0.511	0.599	0.595
ХГ	АЗ	14	**0.927**	**0.928**	0.884	0.86
ХГ	ВД	12	**0.712**	**0.713**	0.699	0.692
ХГ	ГБ	22	0.553	0.553	0.585	0.578
ХГ	ЖК	33	**0.634**	**0.632**	0.58	0.572
ХГ	ИБ	25	0.577	0.577	0.598	0.595
ХГ	МК	41	**0.554**	**0.554**	0.55	0.55
ХГ	ММ	39	**0.639**	**0.639**	0.622	0.611
ХГ	СД	35	0.606	0.606	0.633	0.623

ХГ	УЩ	35	**0.6**	0.6	0.597	0.604
ХХ	АЗ	12	**0.926**	**0.925**	0.897	0.889
ХХ	ВД	26	**0.701**	**0.701**	0.661	0.66
ХХ	ГБ	23	0.578	0.578	0.589	0.588
ХХ	ЖК	35	**0.579**	**0.58**	0.564	0.567
ХХ	ИБ	9	0.572	0.572	0.605	0.597
ХХ	МК	40	0.529	0.53	0.564	0.565
ХХ	ММ	42	0.536	0.536	0.547	0.545
ХХ	СД	29	0.56	0.56	0.594	0.593
ХХ	УЩ	43	**0.567**	**0.566**	0.53	0.539
ХХ	ХГ	31	0.524	0.525	0.557	0.556
ЭА	АЗ	18	**0.878**	**0.877**	0.833	0.819
ЭА	ВД	26	**0.618**	**0.618**	0.596	0.587
ЭА	ГБ	16	**0.69**	**0.69**	0.672	0.655
ЭА	ЖК	16	**0.683**	**0.683**	0.653	0.644
ЭА	ИБ	15	0.701	0.702	0.724	0.715
ЭА	МК	22	0.572	**0.574**	0.581	0.571
ЭА	ММ	27	0.57	0.57	0.601	0.597
ЭА	СД	16	**0.684**	**0.684**	0.624	0.612
ЭА	УЩ	20	**0.59**	**0.59**	0.579	0.57
ЭА	ХГ	13	**0.683**	**0.682**	0.62	0.613
ЭА	ХХ	26	0.594	0.593	0.638	0.632
АП	АЗ	6	0.924	**0.924**	0.927	0.919
АП	ВД	31	**0.789**	**0.789**	0.772	0.762
АП	ГБ	21	0.551	0.55	0.561	0.556
АП	ЖК	36	**0.588**	**0.591**	0.559	0.551
АП	ИБ	12	0.57	0.569	0.578	0.573
АП	МК	37	**0.612**	**0.609**	0.579	0.572
АП	ММ	33	0.631	0.631	0.635	0.636
АП	СД	36	0.565	0.569	0.577	0.569
АП	УЩ	24	**0.632**	**0.63**	0.628	0.625
АП	ХГ	24	0.561	0.562	0.569	0.571
АП	ХХ	29	0.559	0.559	0.567	0.569
АП	ЭА	13	**0.68**	**0.679**	0.656	0.647
АХ	АЗ	15	**0.911**	**0.91**	0.869	0.834
АХ	ВД	29	**0.607**	**0.604**	0.58	0.578
АХ	ГБ	27	0.689	**0.688**	0.689	0.686
АХ	ЖК	21	0.675	0.673	0.727	0.704
АХ	ИБ	17	0.749	0.749	0.793	0.783
АХ	МК	28	**0.694**	**0.694**	0.649	0.636
АХ	ММ	15	0.604	**0.603**	0.612	0.595
АХ	СД	21	0.711	**0.712**	0.714	0.697
АХ	УЩ	24	0.641	0.639	0.658	0.649
АХ	ХГ	16	**0.686**	**0.687**	0.641	0.636
АХ	ХХ	26	**0.694**	**0.693**	0.632	0.628
АХ	ЭА	20	**0.588**	**0.588**	0.572	0.562
АХ	АП	25	0.71	0.709	0.755	0.744
ЯБ	АЗ	10	**0.931**	**0.931**	0.901	0.882

ЯБ	ВД	20	**0.684**	**0.683**	0.652	0.645
ЯБ	ГБ	36	0.616	0.616	0.625	0.629
ЯБ	ЖК	14	0.526	0.525	0.534	0.535
ЯБ	ИБ	33	0.65	0.65	0.663	0.661
ЯБ	МК	30	0.56	0.56	0.567	0.564
ЯБ	ММ	24	0.533	0.533	0.578	0.574
ЯБ	СД	12	0.603	0.602	0.647	0.635
ЯБ	УЩ	21	0.525	0.525	0.563	0.562
ЯБ	ХГ	30	**0.65**	**0.651**	0.584	0.578
ЯБ	ХХ	39	**0.559**	**0.561**	0.54	0.551
ЯБ	ЭА	20	0.568	0.568	0.601	0.597
ЯБ	АП	35	**0.641**	**0.64**	0.576	0.573
ЯБ	АХ	19	0.65	**0.653**	0.661	0.65

4.6 Основные выводы главы 4

1. Одновременный отбор объектов и признаков, осуществляемый алгоритмом mon, частично решает проблему «застревания» в локальных минимумах, связанных с шумовыми объектами и неинформативными признаками.
2. У использования известных функционалов, характеризующих монотонность выборки, в качестве целевых функций алгоритма монотонизации mon, имеется ряд недостатков. Это обосновывает построение специального функционала, характеризующего монотонность выборки, предназначенного специально для использования алгоритмом mon.
3. Параметры предложенного функционала позволяют учитывать несбалансированность классов и различную цену ошибок первого и второго рода.
4. Результаты эксперимента показали, что использование алгоритма предварительной монотонизации mon в ряде случаев позволяет повысить качество классификации.

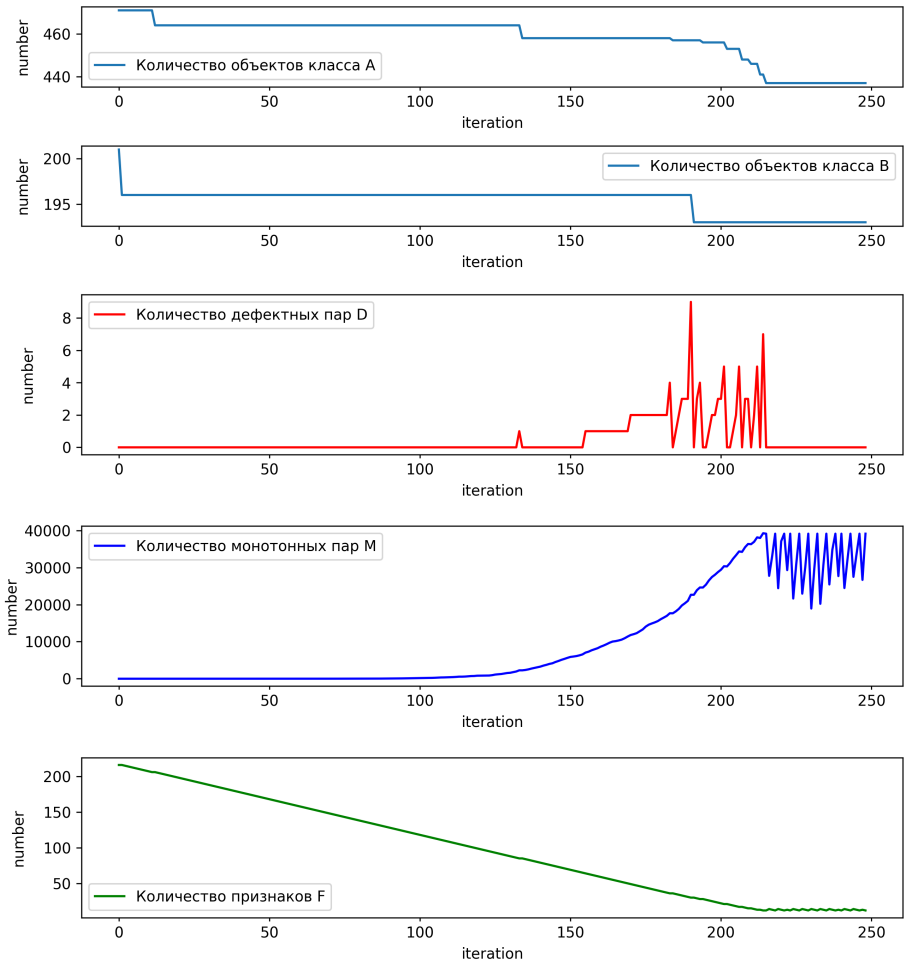


Рисунок 4.4 — Зависимость числа объектов каждого из классов, дефектных пар, монотонных пар а также числа признаков в обучающей подвыборке от номера итерации алгоритма *mon*.

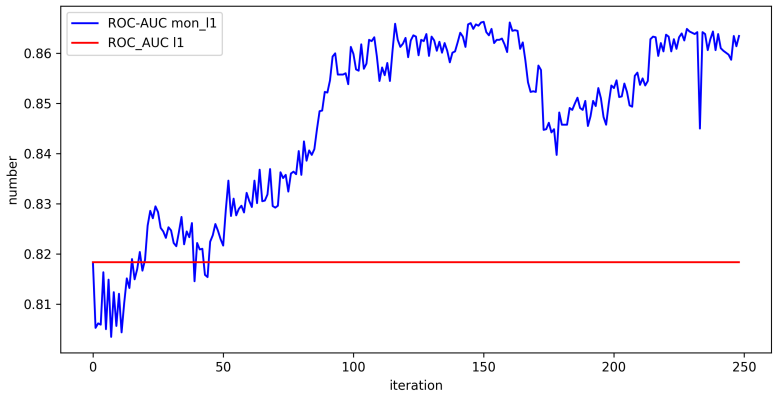


Рисунок 4.5 — Значения ROC-AUC на контрольной подвыборке в зависимости от количества итераций алгоритма *mon* на обучающей подвыборке.

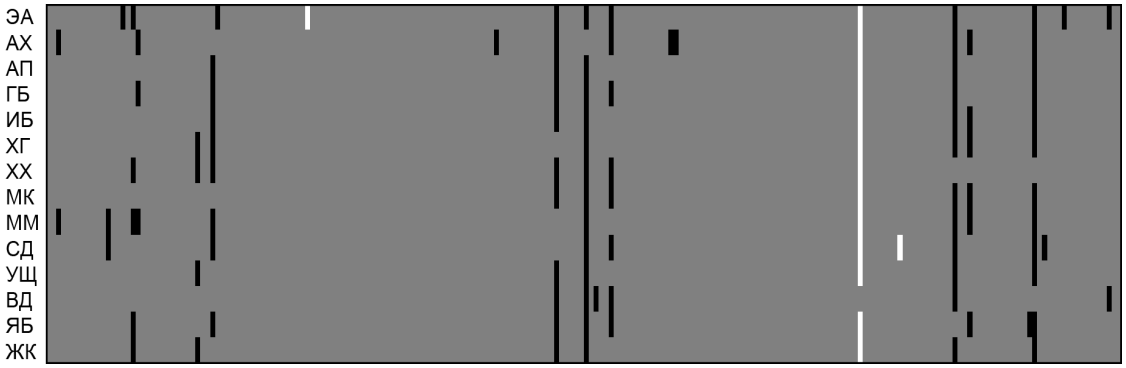


Рисунок 4.6 — Класс АЗ — абсолютно здоров

Рисунок 4.7 — Класс ЭА — анемия железодифицитная

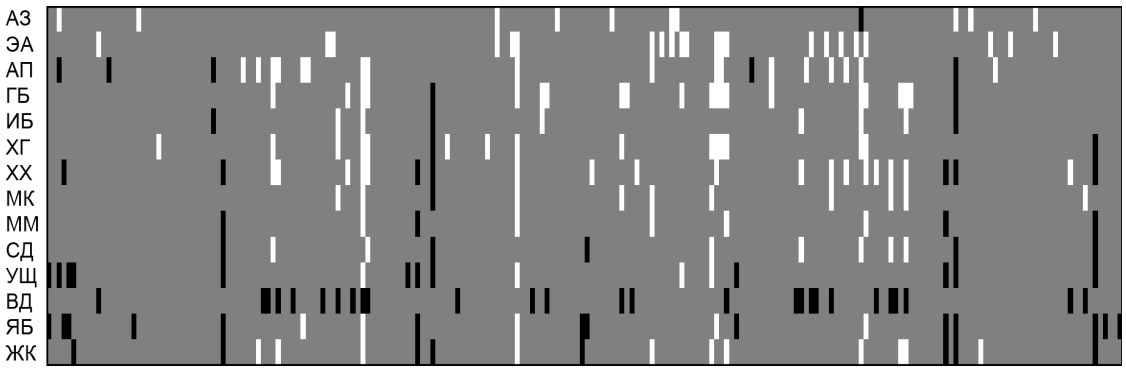


Рисунок 4.8 — Класс АХ — аднексит хронический

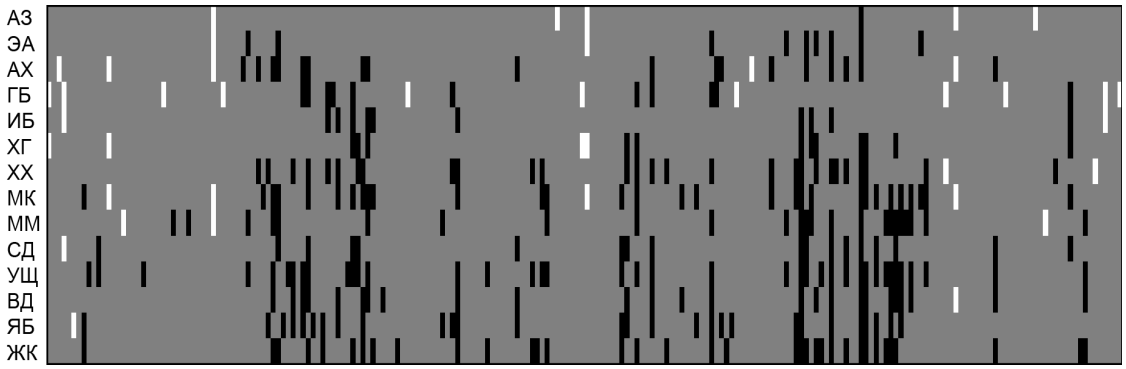


Рисунок 4.9 — Класс АП — аденома простаты

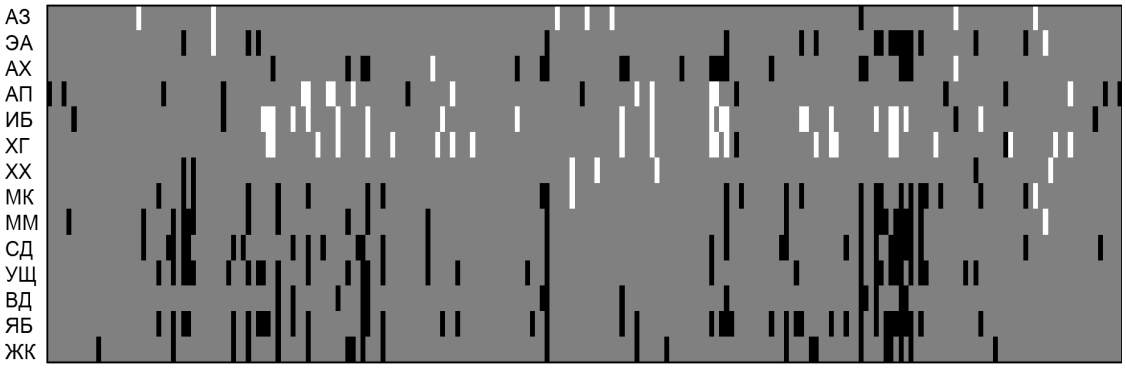


Рисунок 4.10 — Класс ГБ — гипертоническая болезнь

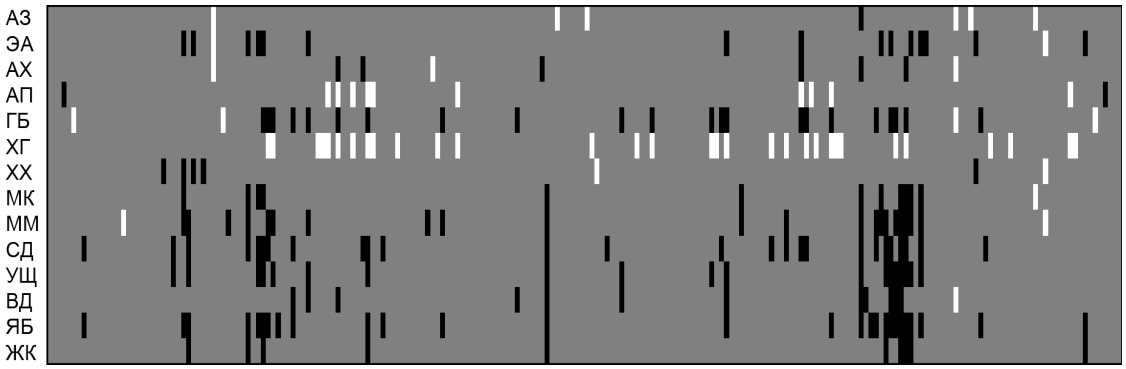


Рисунок 4.11 — Класс ИБ — ишемическая болезнь сердца

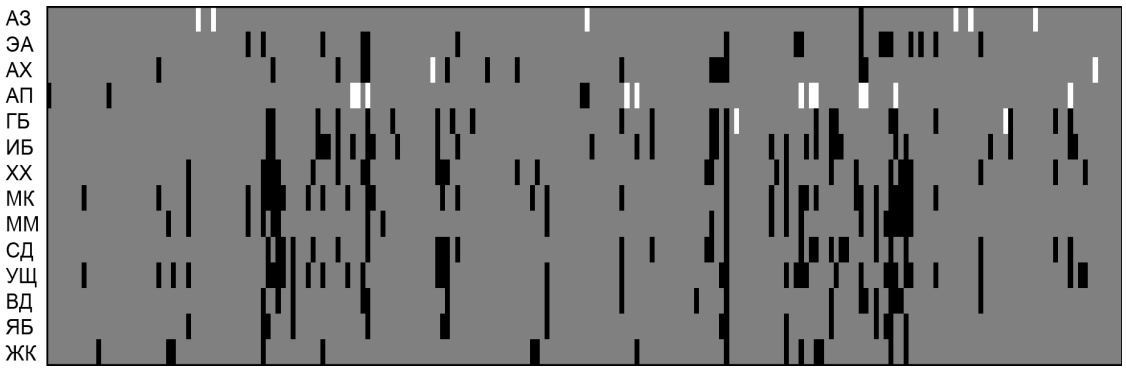


Рисунок 4.12 — Класс ХГ — хронический гастрит гипоацидный

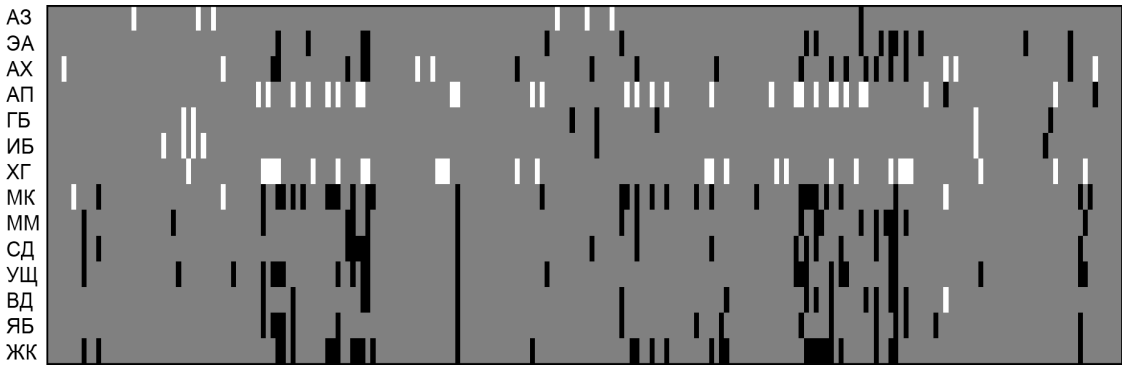


Рисунок 4.13 — Класс ХХ — холицистит хронический

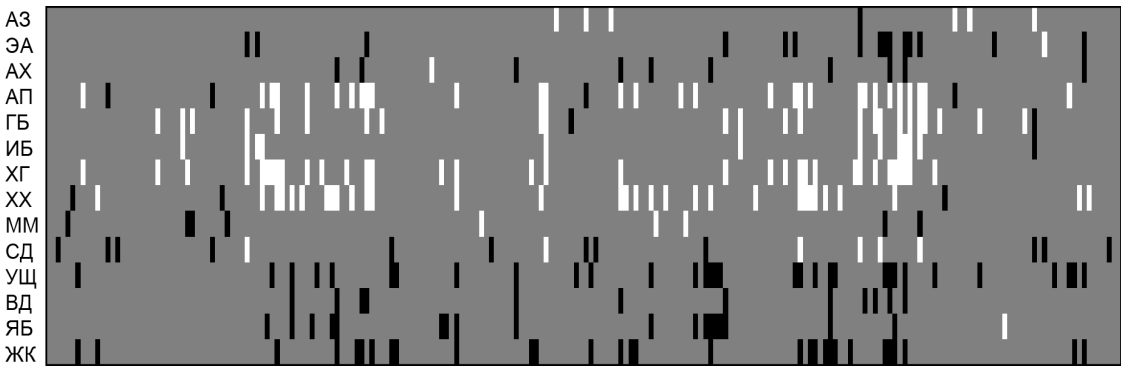


Рисунок 4.14 — Класс МК — мочекаменная болезнь

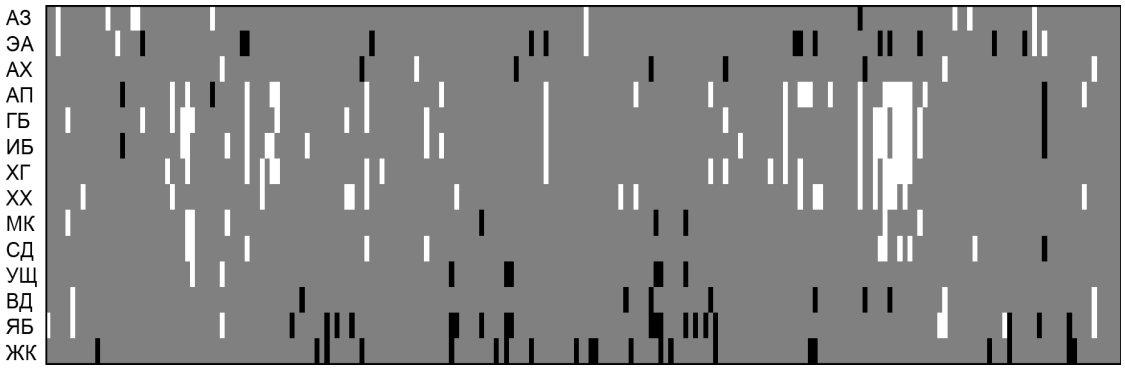


Рисунок 4.15 — Класс ММ — миома матки

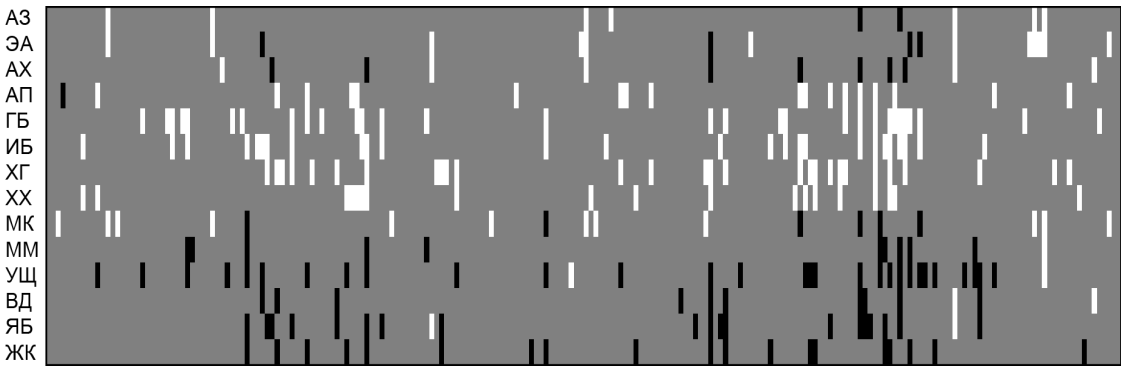


Рисунок 4.16 — Класс СД — сахарный диабет

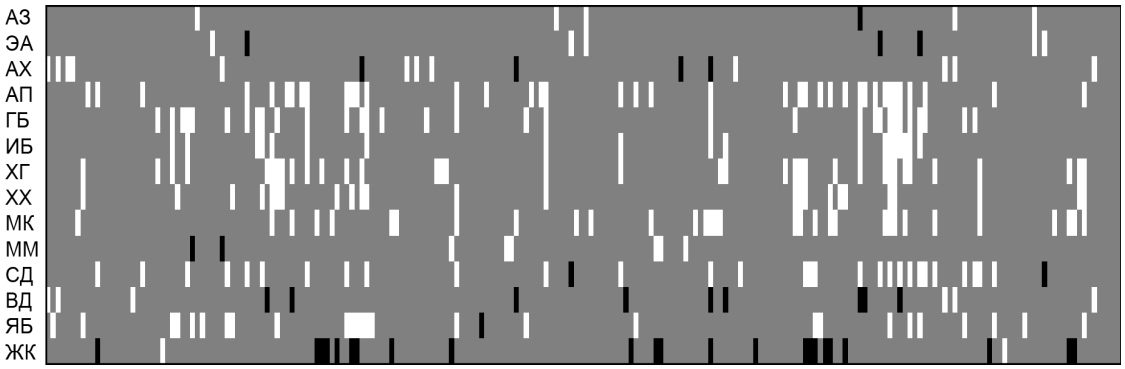


Рисунок 4.17 — Класс УЩ — узловой зоб щитовидной железы

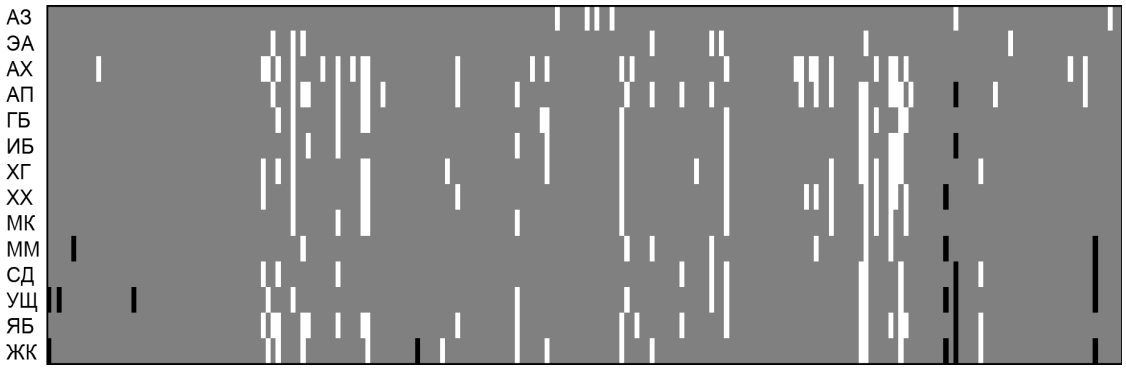


Рисунок 4.18 — Класс ВД — вегетососудистая дистония

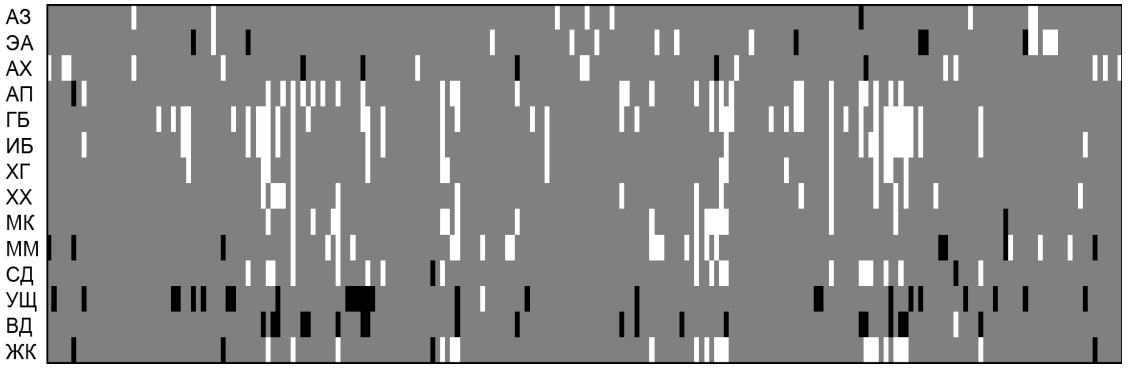


Рисунок 4.19 — Класс ЯБ — язвенная болезнь

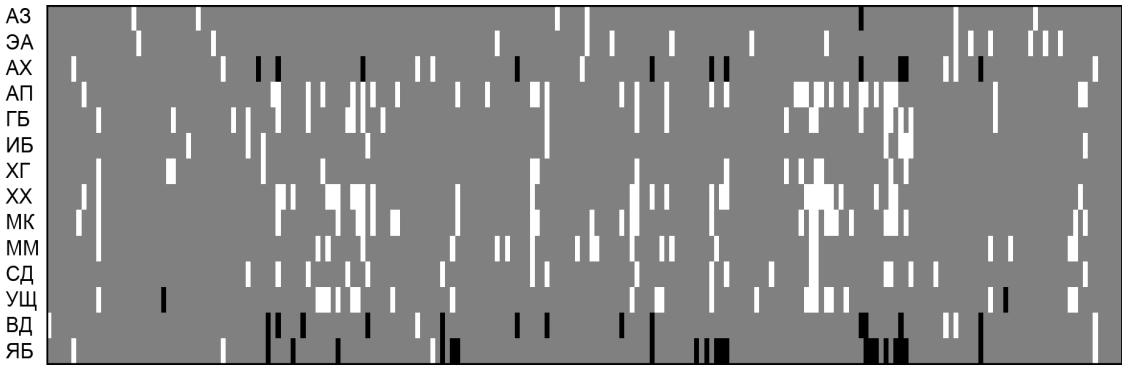


Рисунок 4.20 — Класс ЖК — желчекаменная болезнь

Заключение

Основные результаты работы заключаются в следующем.

1. Получены оценки вычислительной сложности задач отбора объектов и признаков для алгоритма ближайшего соседа. В предложенных постановках данные задачи являются NP-трудными.
2. Получены оценки вычислительной сложности задач отбора объектов и признаков для монотонизации обучающей выборки, предложена их систематизация. Почти все задачи являются NP-трудными, что обосновывает применение приближенных алгоритмов для их решения. Для единственной полиномиальной постановки указан точный эффективный алгоритм решения.
3. Предложен и протестирован экспериментально алгоритм монотонизации выборки с одновременным отбором объектов и признаков. Показано, что в ряде случаев использование предложенного алгоритма повышает качество классификации.

Возможным направлением дальнейших исследований является изучение задач отбора объектов и признаков для более широкого семейства классификаторов и разработка эффективных алгоритмов для их решения.

Список сокращений и условных обозначений

Y	множество ответов
(x_i,y_i)	прецедент
X^L	обучающая выборка
X^ℓ	обучающая подвыборка
X^k	контрольная подвыборка
γ	алгоритм
Ω	множество эталонных объектов
$I(x,\gamma(x))$	индикатор ошибки алгоритма γ на объекте x
$\nu(\gamma,X)$	частота ошибок алгоритма γ на выборке X
$Q_k(\mu)$	функционал полного скользящего контроля с контрольной выборкой длинны k
$\rho(x,x')$	функция расстояния
$r_m(x_i)$	ошибка, возникающая при замене известной классификации объекта x_i на ответ $y(x_{im})$ на m-ом соседе
$P(m)$	профиль компактности
$\mathbb{F}$	множество признаков
$F \subseteq \mathbb{F}$	подмножество признаков
$\mathbb{A}$	множество объектов обучающей выборки класса 1
$\mathbb{B}$	множество объектов обучающей выборки класса 0
M	множество монотонных пар
D	множество дефектных пар
M_f	множество пар, монотонных по признаку f
D_f	множество пар, дефектных по признаку f
M_F	множество пар, монотонных совокупности признаков F
D_F	множество пар, дефектных совокупности признаков F
Ψ^F	матрица, описывающая частичный порядок, индуцируемый множеством признаков F на объектах обучающей выборки

$\overline{\Psi}$	трехмерное представление, описывающее частичный порядок, индуцируемый множеством признаков F на объектах обучающей выборки с учетом инверсий признаков

Список литературы

1. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.

2. *Журавлев Ю.И., Рязанов В.В., Сенько О.В.* Математические методы. Программная система. Практические применения. — М.: Фазис, 2005. — 159 с.

3. *Кельманов А.В.* О сложности некоторых задач анализа данных // *Журнал вычислительной математики и математической физики.* — 2010. — Т. 50, № 11. — С. 2045–2051.

4. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999. — 270 с.

5. *Bermejo S., Cabestany J.* Learning with nearest neighbour classifiers // *Neural Processing Letters.* — 2001. — Vol. 13, №.2. – P. 159–181.

6. Сходство и компактность / И.А. Борисова, В.В. Дюбанов, Н.Г. Загоруйко, О.А. Кутненко // Математические методы распознавания образов: 14-я Всероссийская конференция, г.Суздаль, 21–26 сентября 2009 г.: Сборник докладов. — М.: МАКС Пресс, 2009. — С. 89–92.

7. *Воронцов К.В., Колосков А.О.* Профили компактности и выделение опорных объектов в метрических алгоритмах классификации // *Искусственный Интеллект.* — 2006. — № 2. — С. 30–33.

8. *Иванов М.Н., Воронцов К.В.* Отбор эталонов, основанный на минимизации функционала полного скользящего контроля // Математические методы распознавания образов: 14-я Всероссийская конференция, г.Суздаль, 21–26 сентября 2009 г.: Сборник докладов. — М.: МАКС Пресс, 2009. — С. 119–122.

9. *Sill J., Abu-Mostafa Y.S.* Monotonicity hints // *Advances in Neural Information Processing Systems – Cambridge: MIT Press, 1997.* — Vol. 9. – P. 634–640.

10. *Sill J.* Monotonic networks // *Advances in Neural Information Processing Systems – Cambridge: MIT Press, 1998.* — Vol. 10. – P. 661–667.

11. *Воронцов К.В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // *Математические вопросы кибернетики / Под ред. Лупанова О.Б. – М.:Физматлит, 2004.* — Т. 13. — С. 5–36.

12. *Malar B., Nadarajan R., Saisundarakrishnan G.* Isotonic separation with an instance selection algorithm using softset: Theory and experiments // *WSEAS Transactions On Information Science And Applications.* — 2012. — Vol. 9, №.11. – P. 350–367.

13. *Royston P.* A useful monotonic non-linear model with applications in medicine and epidemiology // *Statistics in Medicine.* — 2000. — Vol. 19, №.15. – P. 2053–2066.

14. *Ryu Y.U., Chandrasekaran R., Jacob V.S.* Breast cancer prediction using the isotonic separation technique // *European Journal of Operational Research.* — 2007. — Vol. 181, №.2. – P. 842–854.

15. *Spirin N.V., Vorontsov K.V.* Learning to rank with nonlinear monotonic ensemble // 10th International Workshop on Multiple Classifier Systems (Naples, Italy, June 15–17, 2011): Lecture Notes in Computer Science / Ed. by Roli F. Sansone C., Kittler J. — Berlin: Springer-Verlag, 2011. – P. 16–25.

16. *Иванов М.Н. and Воронцов К.В.* Применение монотонного классификатора ближайшего соседа в задаче категоризации текстов // Интеллектуализация обработки информации ИОИ: 9-я международная конференция. Черногория, г. Будва, 2012 г.: Сборник докладов. — М.: ТОРУС ПРЕСС, 2012. — С. 621–624.

17. *Гуз И.С.* Нелинейные монотонные композиции классификаторов // Математические методы распознавания образов: 13-я Всероссийская конференция, г.Зеленогорск, 30 сентября – 6 октября 2007 г.: Сборник докладов. — М.: МАКС Пресс, 2007. — С. 111–114.

18. *Гуз И.С.* Минимизация эмпирического риска при построении монотонных композиций классификаторов // *Труды МФТИ.* — 2011. — Т. 3, № 3(11). — С. 115–121.

19. Isotonic separation / R. Chandrasekaran, Y.U. Ryu, V.S. Jacob, S. Hong // *INFORMS Journal on Computing.* — 2005. — Vol. 17, №.4. – P. 462–474.

20. *Kamp R., Feelders A., Barile N.* Isotonic classification trees // Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII. — Berlin, Heidelberg: Springer-Verlag, 2009. – P. 405–416.

21. *Marsala C., Petturiti D.* Monotone Classification with Decision Trees // 8th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2013). — Atlantis Press, 2013. – P. 810–817.

22. *Воронцов К.В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // *ЖВМ и МФ.* — 2000. — Т. 40, № 1. — С. 166–176.

23. *Воронцов К.В., Махина Г.А.* Принцип максимизации зазора для монотонного классификатора ближайшего соседа // Математические методы распознавания образов: 15-я Всероссийская конференция, г.Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011. — С. 64–67.

24. *Махина Г.А.* О восстановлении монотонных булевых функций методом ближайшего соседа // Интеллектуализация обработки информации ИОИ: 9-я международная конференция. Черногория, г. Будва, 2012 г.: Сборник докладов. — М.: ТОРУС ПРЕСС, 2012. — С. 78–81.

25. *Cortes C., Vapnik V.* Support-vector networks // *Machine Learning.* — 1995. — Vol. 20, №.3. – P. 273–297.

26. *Воронцов К.В.* О проблемно-ориентированной оптимизации базисов задач распознавания // *ЖВМ и МФ.* — 1998. — Т. 38, № 5. — С. 870–880.

27. Statistical inference under order restrictions; the theory and application of isotonic regression / R. Barlow, D. Bartholomew, J. Bremner, H. Brunk. — New York: Wiley, 1972.

28. *Feelders A., Velikova M., Daniels H.* Tech. Rep. UU-CS-2006-046: Two polynomial algorithms for relabeling non-monotone data. — Department of Information and Computing Sciences, Utrecht University: 2006.

29. Statistical approach to ordinal classification with monotonicity constraints / W. Kotlowski, R. Slowinski, E. Hullermeier, J. Fiimkranz // ECML PKDD 2008 Workshop on Preference Learning. — 2008.

30. Information function of the heart: Discrete and fuzzy encoding of the ECG-signal for multidisease diagnostic system / V.M. Uspenskiy, K.V. Vorontsov, V.R. Tselykh, V.A. Bunakov // Advanced Mathematical and Computational Tools in Metrology and Testing X (vol. 10) (Series on Advances in Mathematics for Applied Sciences). — Vol. 86. — World Scientific Publishing Company Singapore, 2015. — Pp. 377–384.

31. *Зухба А.В.* NP-полнота задачи оптимального отбора эталонных объектов в методе ближайшего соседа // Труды 52-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». Часть VII. Управление и прикладная математика. Том 2. — М.: МФТИ, 2009. — С. 61–63.

32. *Зухба А.В.* Оценка оптимальности жадного алгоритма отбора эталонных объектов в методе ближайшего соседа // Труды 53-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук». Часть VII. Управление и прикладная математика. Том 2. — М.: МФТИ, 2010. — С. 75–76.

33. *Зухба А.В.* Сложность задачи отбора эталонов в методе ближайшего соседа // Математические методы распознавания образов: 15-я Всероссийская конференция, г.Петрозаводск, 11–17 сентября 2011 г.: Сборник докладов. — М.: МАКС Пресс, 2011. — С. 305–308.

34. *Зухба А.В.* Оценка вычислительной сложности задачи монотонизации выборки // Математические методы распознавания образов: 16-я Всероссийская конференция, г.Казань, 6–12 сентября 2013 г.: Тезисы докладов. — М.: ТОРУС ПРЕСС, 2013. — С. 39.

35. *Зухба А.В.* Отбор объектов и признаков для монотонных классификаторов // Математические методы распознавания образов: Тезисы докладов 17-й Всероссийской конференции с международным участием, г.Светлогорск, 2015 г. — М.: ТОРУС ПРЕСС, 2015. — С. 96–97.

36. *Швец М.Ю. Зухба А.В. Воронцов К.В.* Построение монотонного классификатора для задач медицинской диагностики // Математические методы распознавания образов: Тезисы докладов 17-й Всероссийской конференции с международным участием, г.Светлогорск, 2015 г. — М.: ТОРУС ПРЕСС, 2015. — С. 42–43.

37. *Зухба А.В.* Алгоритм монотонизации выборки с одновременным отбором объектов и признаков // Математические методы распознавания образов: Тезисы докладов 18-й Всероссийской конференции с международным участием, г.Таганрог, 2017 г. — М.: ТОРУС ПРЕСС, 2017. — С. 52.

38. *Zukhba A.V.* NP-Completeness of the Problem of Prototype Selection in the Nearest Neighbor Method // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, №.4. – P. 484–494.

39. *Зухба А.В.* Вычислительная сложность отбора объектов и признаков для задач классификации с ограничениями монотонности [Электронный ресурс] // *Математическая биология и биоинформатика.* — 2015. — Т. 10, № 2. — С. 356–371. — Режим доступа: http://www.matbio.org/article.php?journ_id=22&id=244. – (Дата обращения: 25.01.2018).

40. *Зухба А.В.* Оценка вычислительной сложности задачи монотонизации множества при помощи отбора признаков // *Математические и информационные модели управления: сб. науч. трудов.* — М.: МФТИ, 2013. — С. 124–132.

41. *Vapnik V.N.* Statistical learning theory. — N.Y.: John Wiley and Sons, Inc., 1998.– 732 p.

42. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning, 2nd edition. — Springer, 2009.– 533 p.

43. *Воронцов К.В.* Комбинаторная теория надёжности обучения по прецедентам: Диссертация на соискание ученой степени доктора физико-математических наук: 05.13.17/ Вычислительный центр РАН. — М., 2010. — 230 с.

44. *Mullin M., Sukthankar R.* Complete Cross-Validation for Nearest Neighbor Classifiers // Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 – July 2, 2000. — 2000. – P. 639–646.

45. *Аркадьев А.Г., Браверман Э.М.* Обучение машины распознаванию образов. — М.: Наука, 1964. — 112 с.

46. Алгоритмы: построение и анализ, 3-е издание / Кормен Т., Лейзерсон Ч., Риверсет Р., Штайн К. — М.: Вильямс, 2015. — 1328 с.

47. *Гэри М., Джонсон Д.* Вычислительные машины и труднорешаемые задачи. — М.: Мир, 1982. — 416 с.

48. *Дасгупта С., Пападимитриу Х., Вазирани У.* Алгоритмы. — М.: МЦНМО, 2014. — 320 с.

49. *Корте Б., Фиген Й.* Комбинаторная оптимизация. Теория и алгоритмы. — М.: МЦНМО, 2015. — 720 с.

50. *Ivanov M.N.* Prototype sample selection based on minimization of the complete cross validation functional // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, №.4. – P. 427–437.

51. *Zhang J.* Advancements of Outlier Detection: A Survey // *ICST Transactions on Scalable Information Systems.* — 2013. — Vol. 13, №.1. – P. 1–26.

52. *Velikova M., Daniels H.* On Testing Monotonicity of Datasets // Workshop on Learning Monotone Models from Data at European Conference on Machine Learning and Principles of Knowledge Discovery in Databases, Bled, Slovenia, 2009. – P. 11 – 22.

53. *Успенский В.М.* Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного

анализа электрокардиосигналов. — М.: Экономика и информатика, 2008. — 151 с.

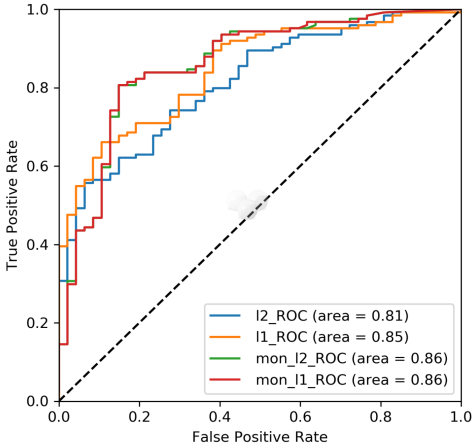
54. *Heagerty P.J., Lumley T., Pepe M.S.* Time-dependent ROC Curves for Censored Survival Data and a Diagnostic Marker // *Biometrics*. — 2000. — Vol. 56, – P. 337–344.

Список рисунков

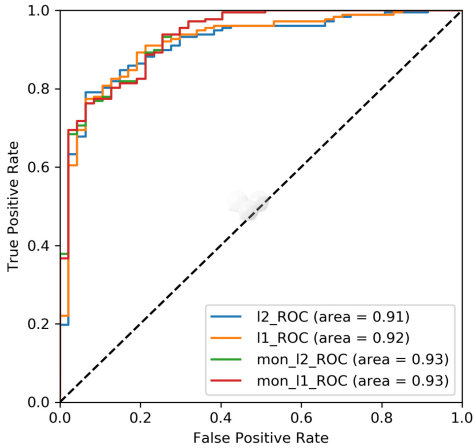
Список таблиц

Приложение А

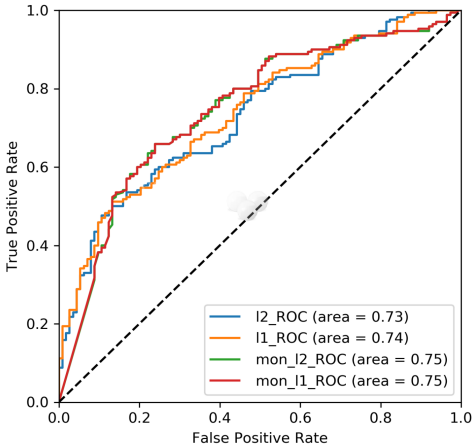
ROC-кривые



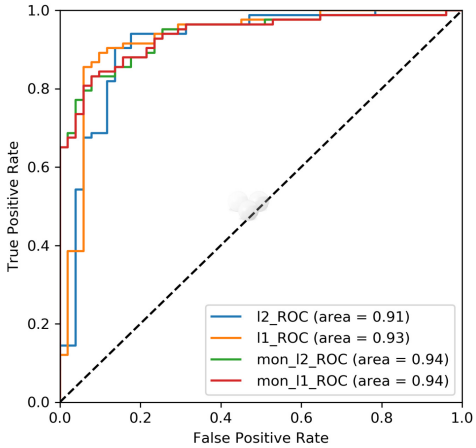
Задача ВД–АЗ



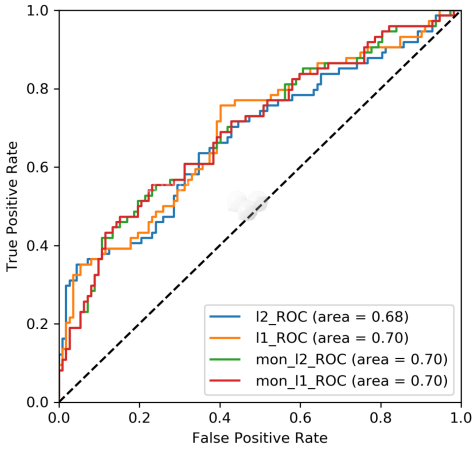
Задача ГБ–АЗ



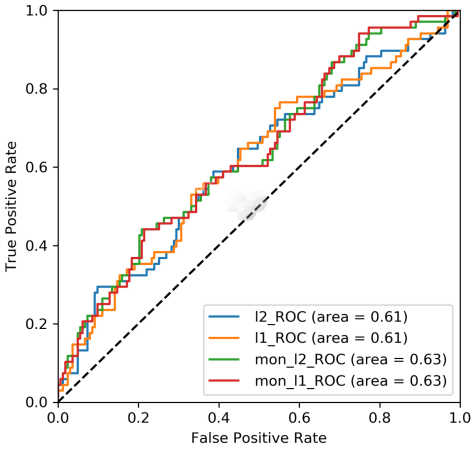
Задача ГБ–ВД



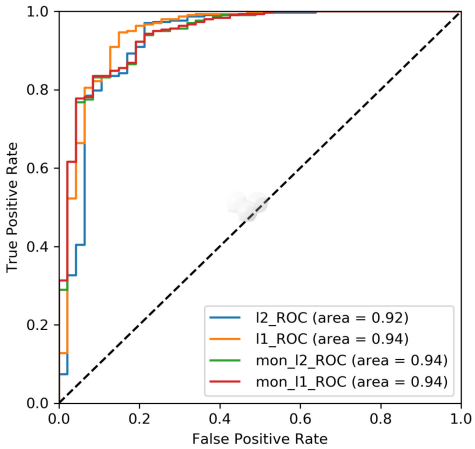
Задача ЖК–АЗ



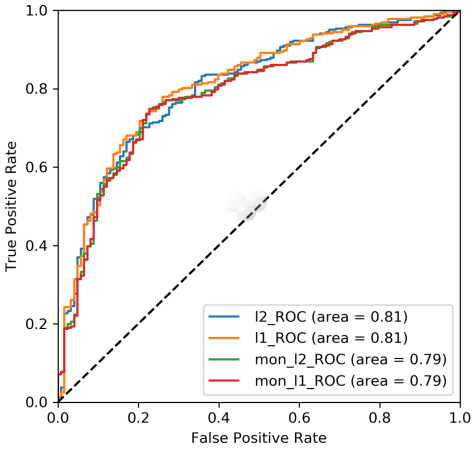
Задача ЖК–ВД



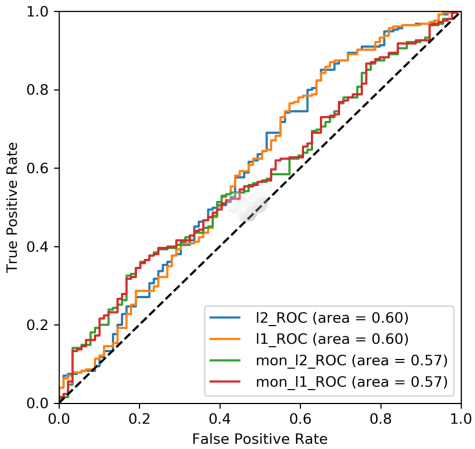
Задача ЖК–ГБ



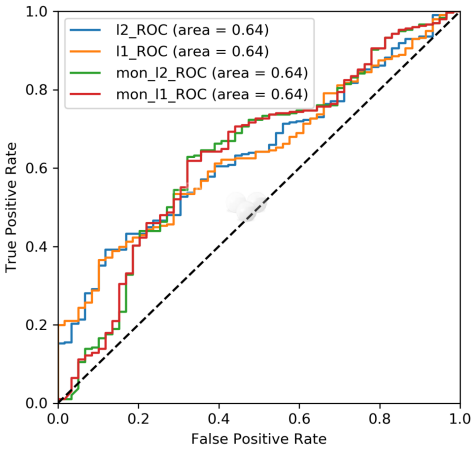
Задача ИБ–АЗ



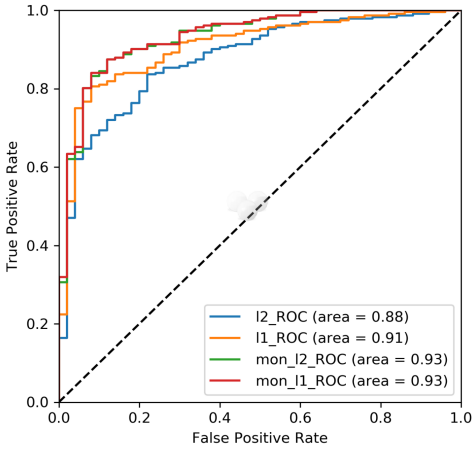
Задача ИБ–ВД



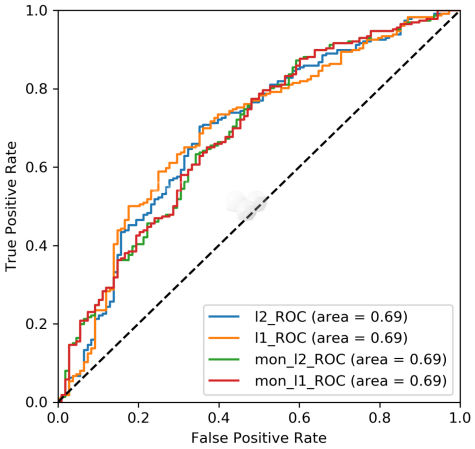
Задача ИБ–ГБ



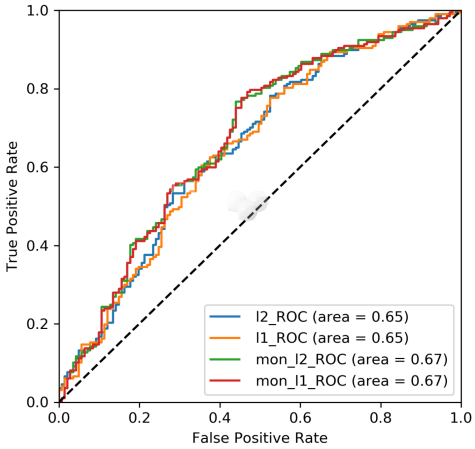
Задача ИБ–ЖК



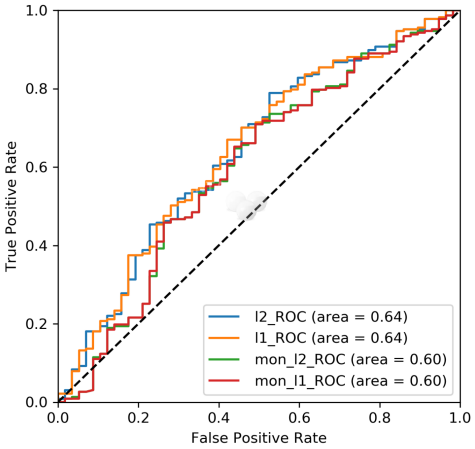
Задача МК–АЗ



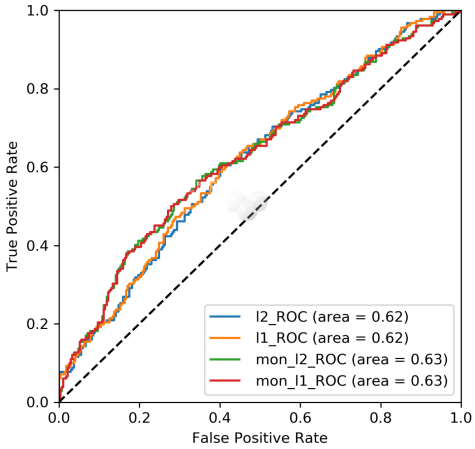
Задача МК–ВД



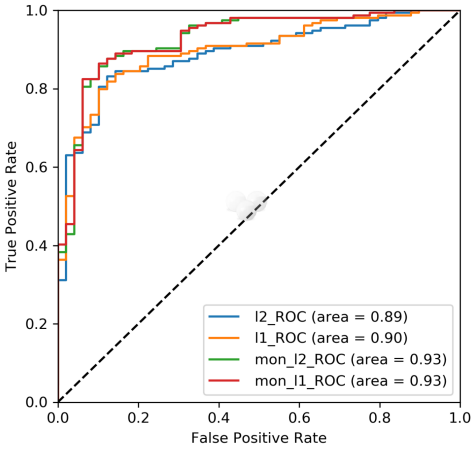
Задача МК–ГБ



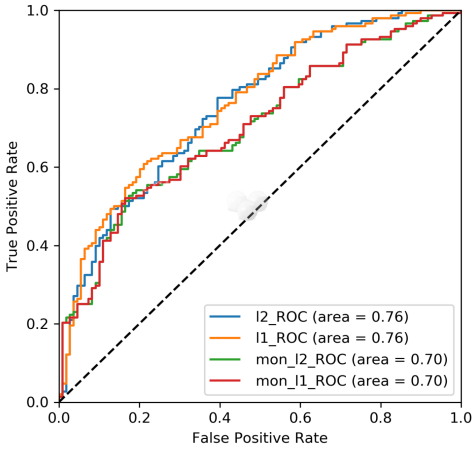
Задача МК–ЖК



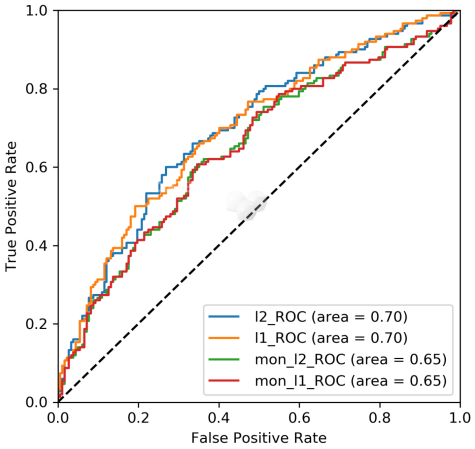
Задача МК–ИБ



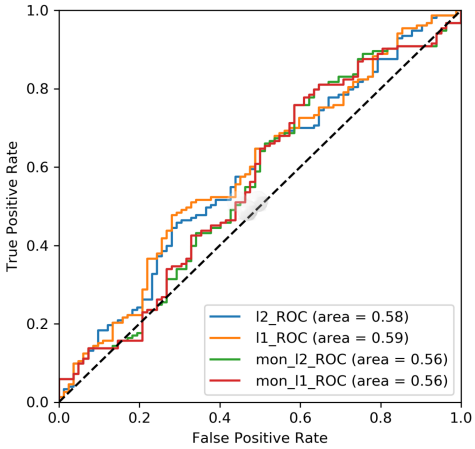
Задача ММ–АЗ



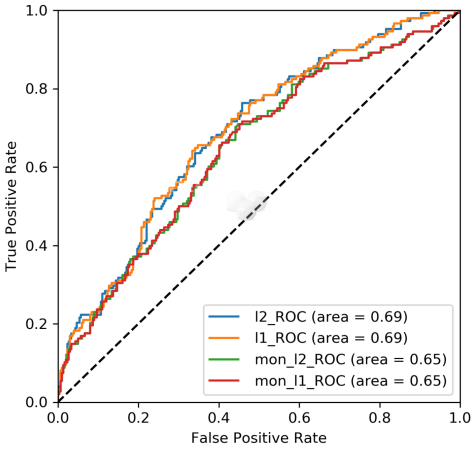
Задача ММ–ВД



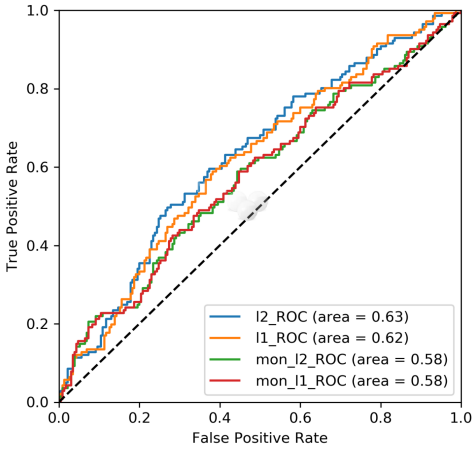
Задача ММ–ГБ



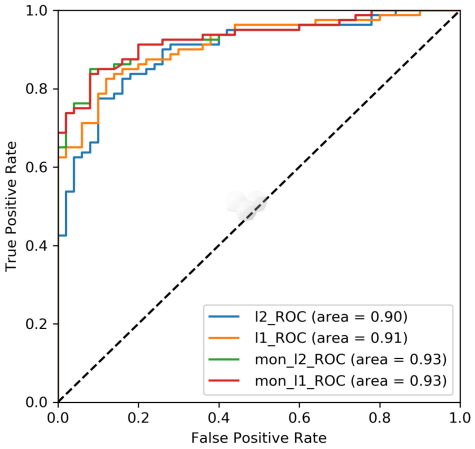
Задача ММ–ЖК



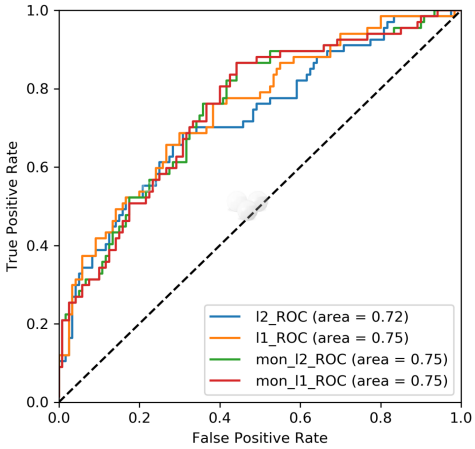
Задача ММ–ИБ



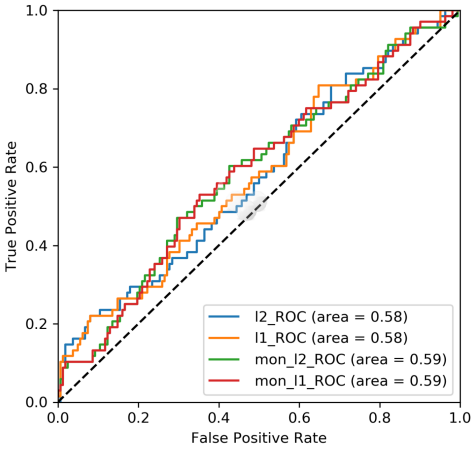
Задача ММ–МК



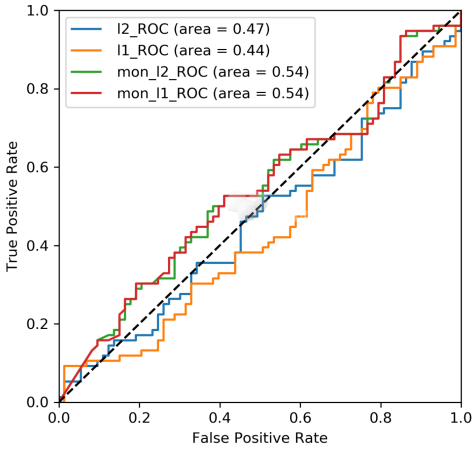
Задача СД–АЗ



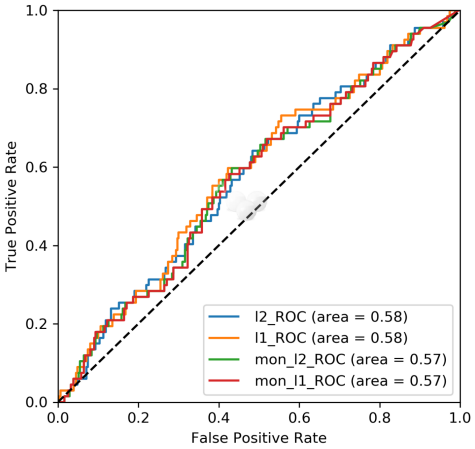
Задача СД–ВД



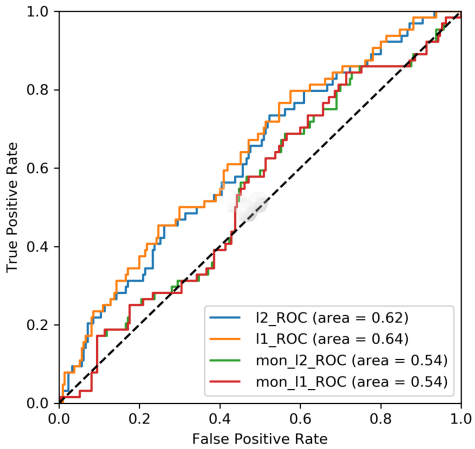
Задача СД–ГБ



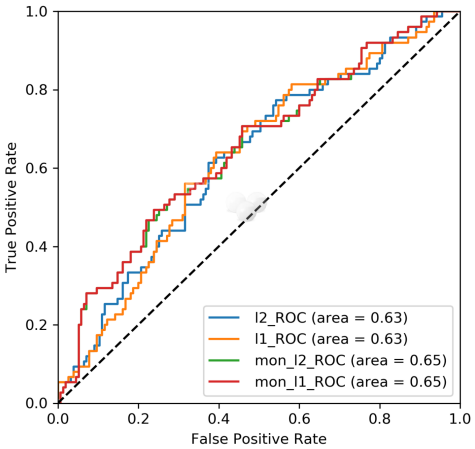
Задача СД–ЖК



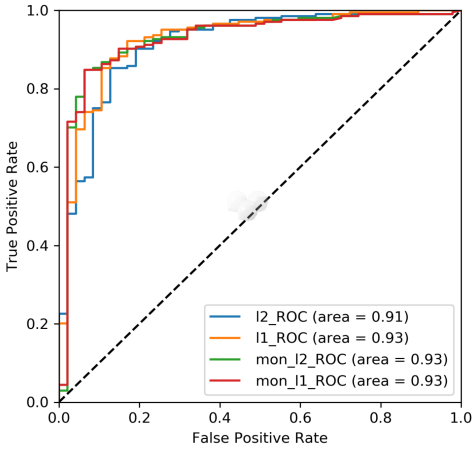
Задача СД–ИБ



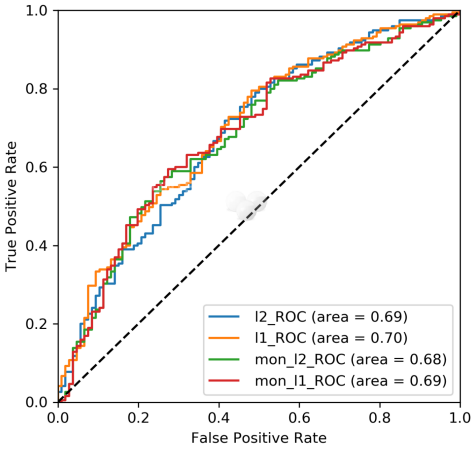
Задача СД–МК



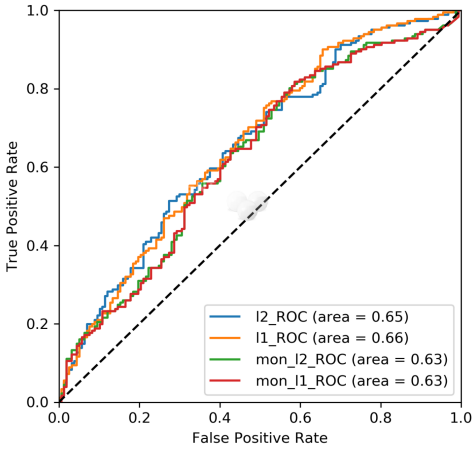
Задача СД–ММ



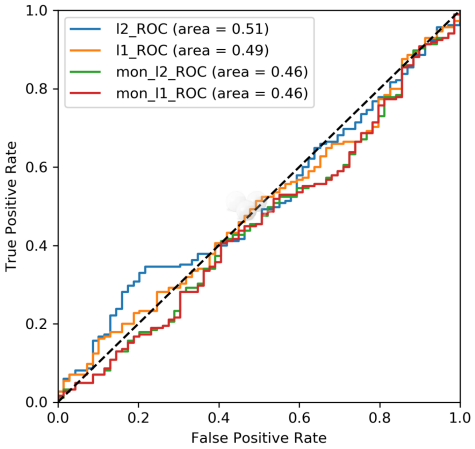
Задача УЩ–АЗ



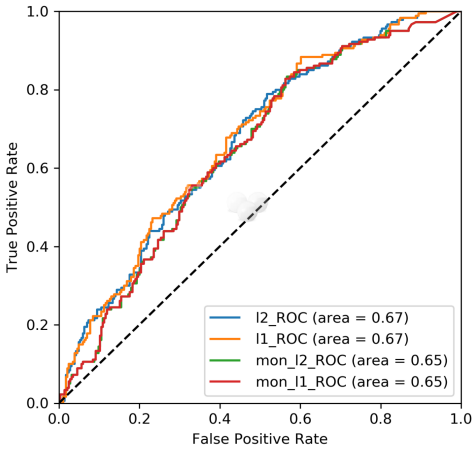
Задача УЩ–ВД



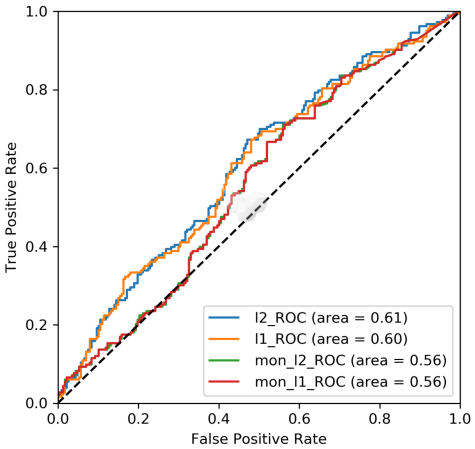
Задача УЩ–ГБ



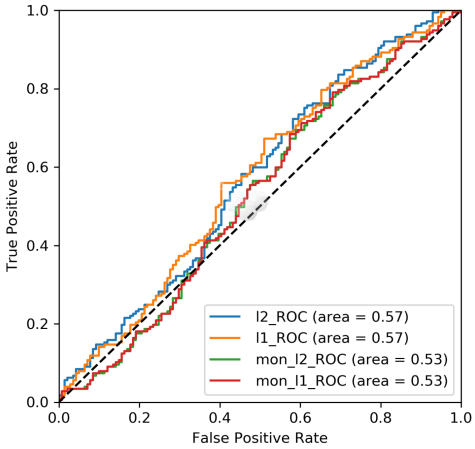
Задача УЩ–ЖК



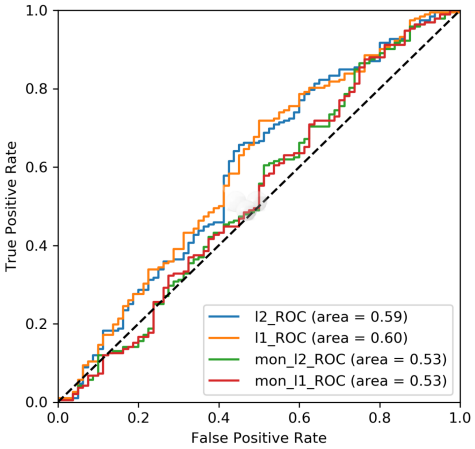
Задача УЩ–ИБ



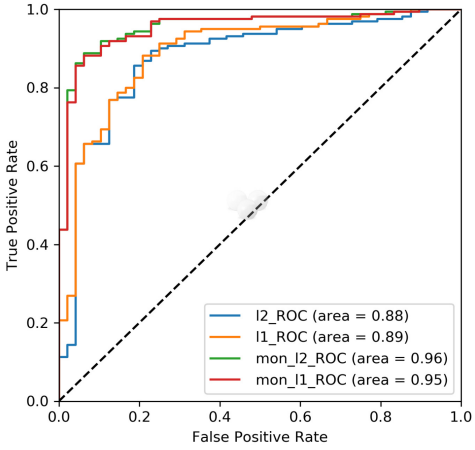
Задача УЩ–МК



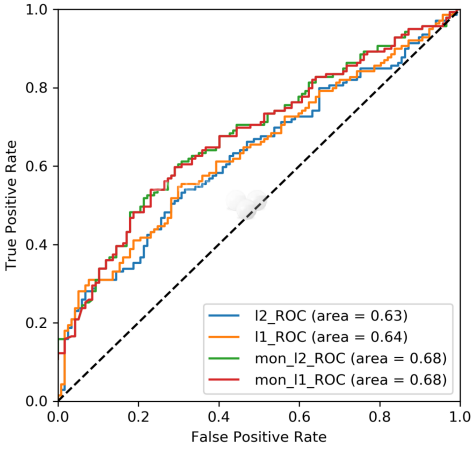
Задача УЩ–ММ



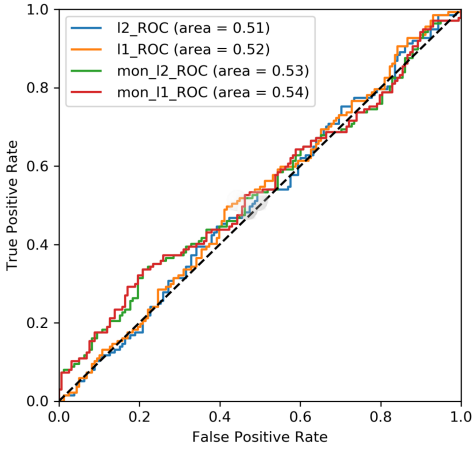
Задача УЩ–СД



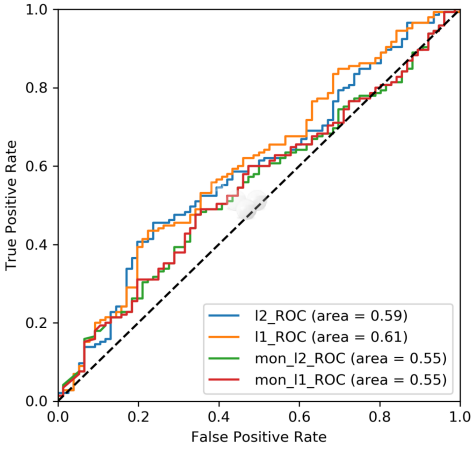
Задача ХГ–АЗ



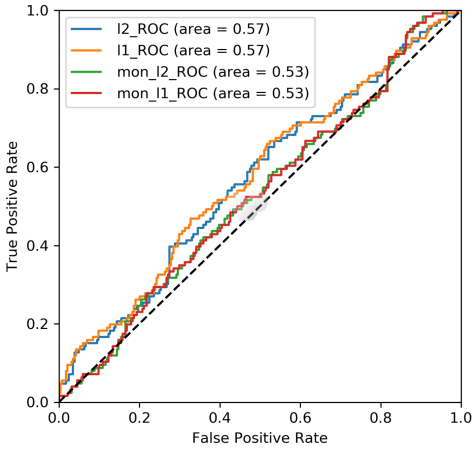
Задача ХГ–ВД



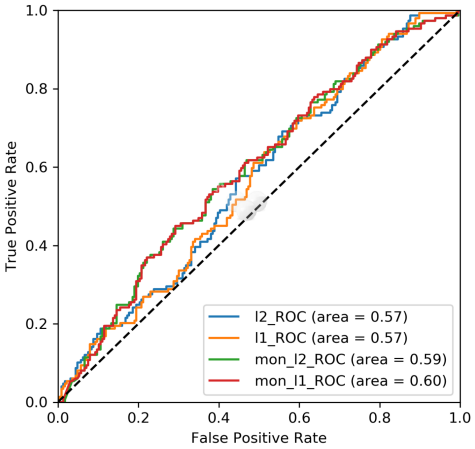
Задача ХГ–ГБ



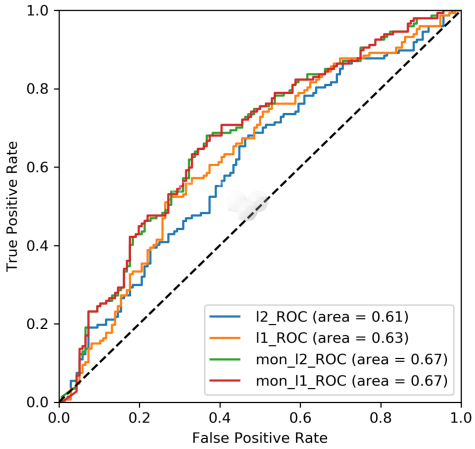
Задача ХГ–ЖК



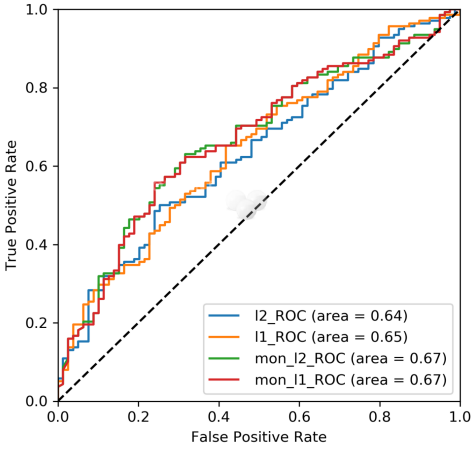
Задача ХГ–ИБ



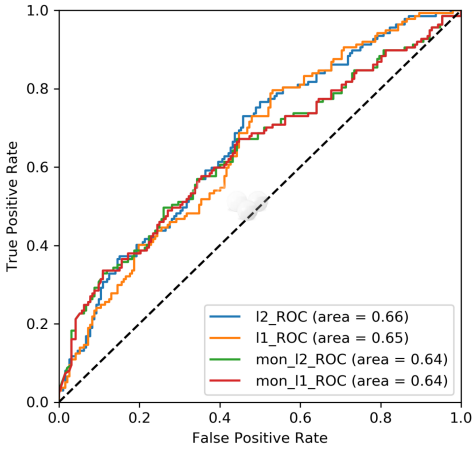
Задача ХГ–МК



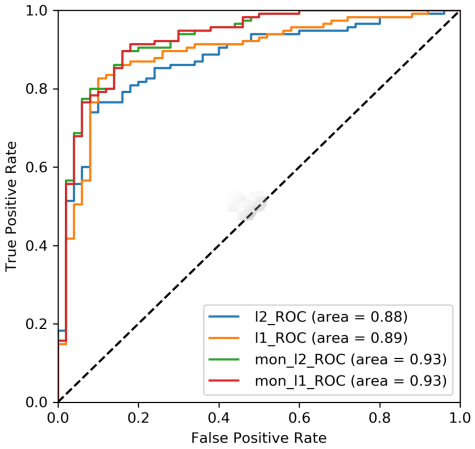
Задача ХГ–ММ



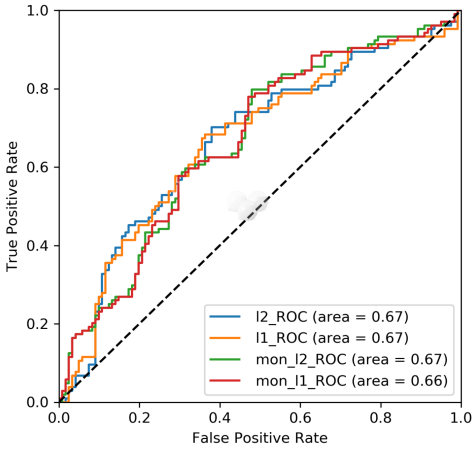
Задача ХГ–СД



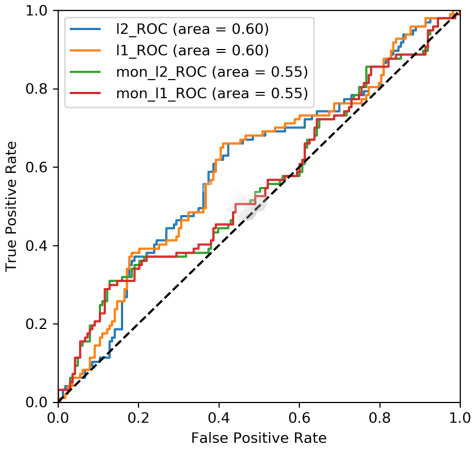
Задача ХГ–УЩ



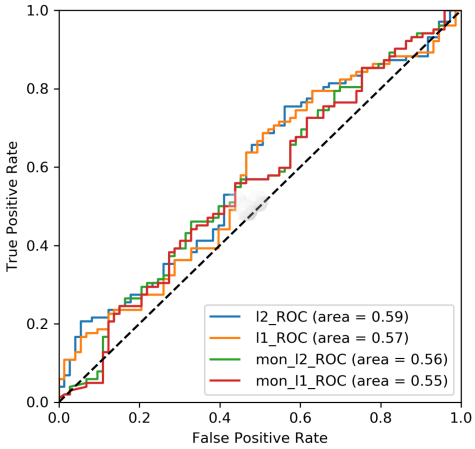
Задача ХХ–АЗ



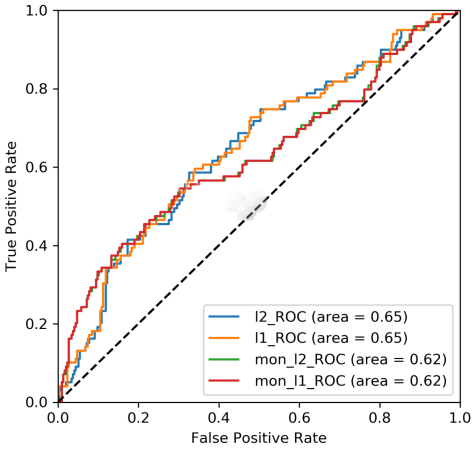
Задача ХХ–ВД



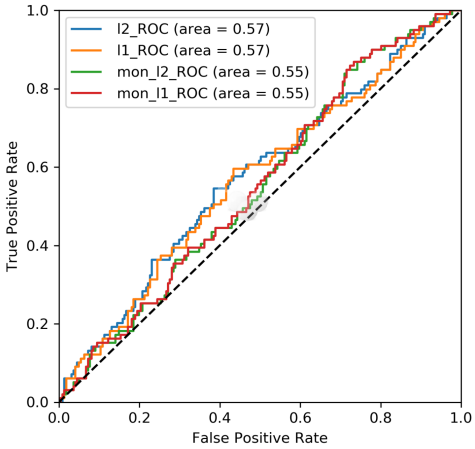
Задача ХХ–ГБ



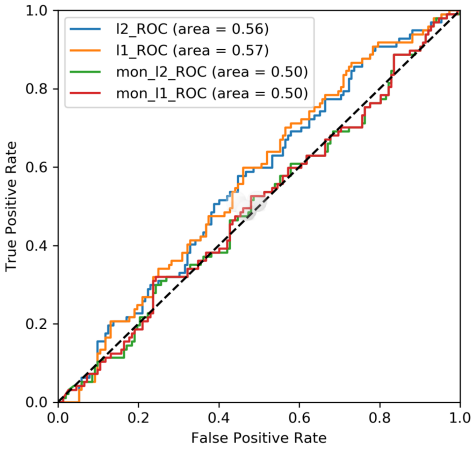
Задача ХХ–ЖК



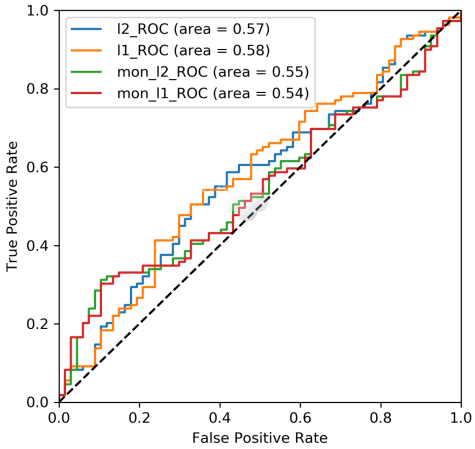
Задача ХХ–ИБ



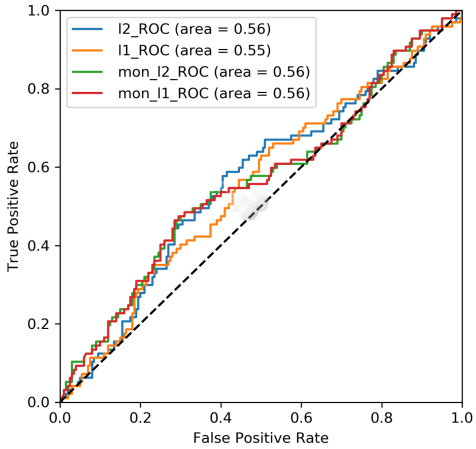
Задача ХХ–МК



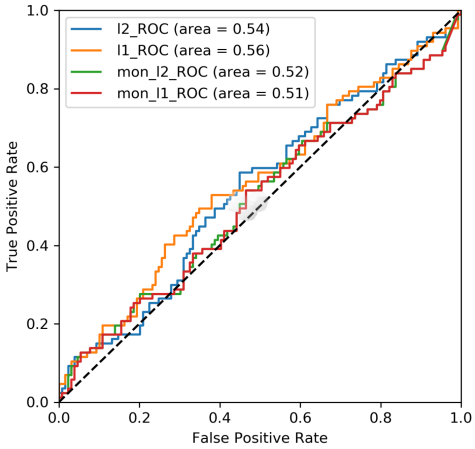
Задача ХХ–ММ



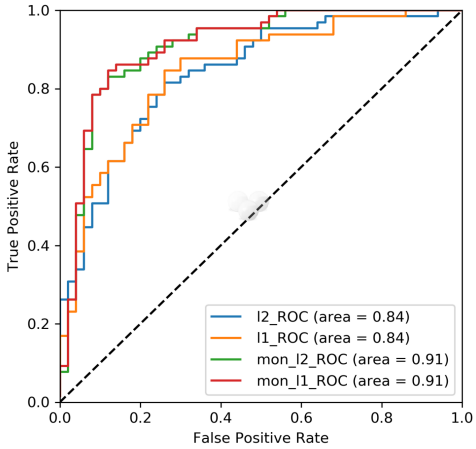
Задача ХХ–СД



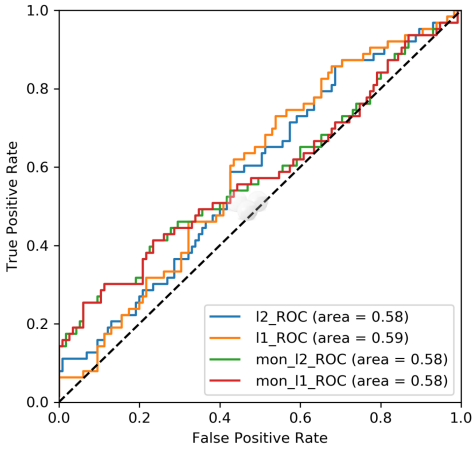
Задача ХХ–УЩ



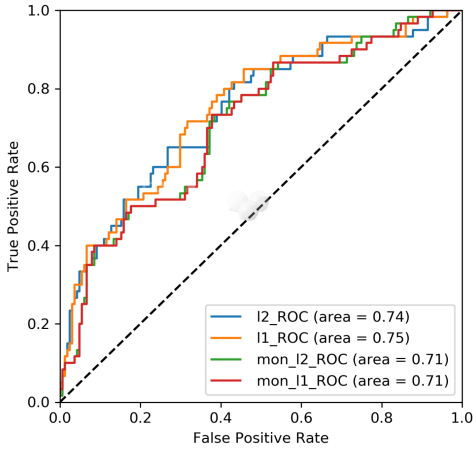
Задача ХХ–ХГ



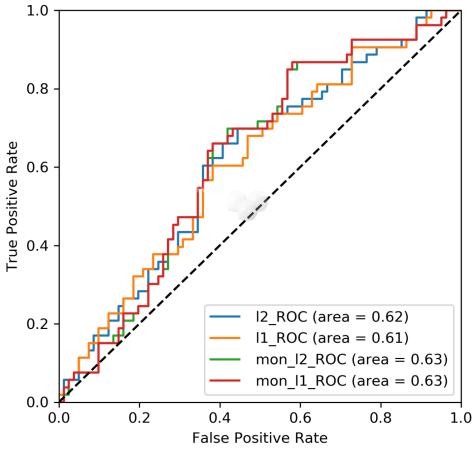
Задача ЭА–АЗ



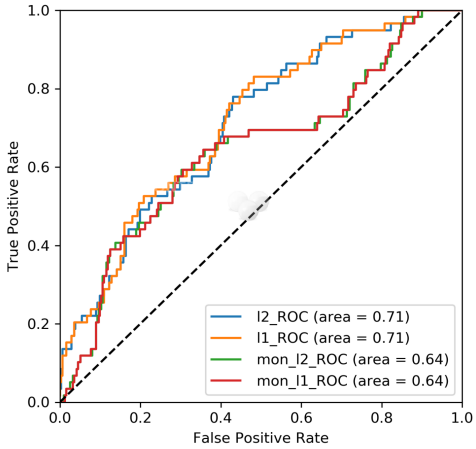
Задача ЭА–ВД



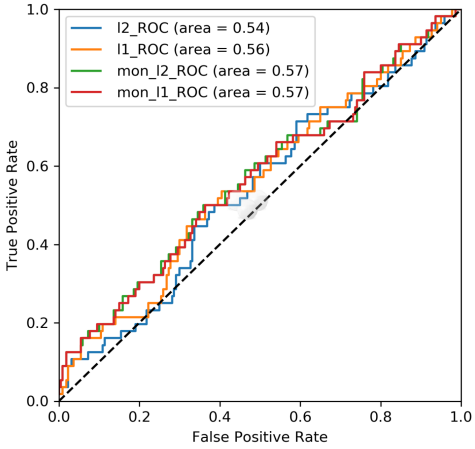
Задача ЭА–ГБ



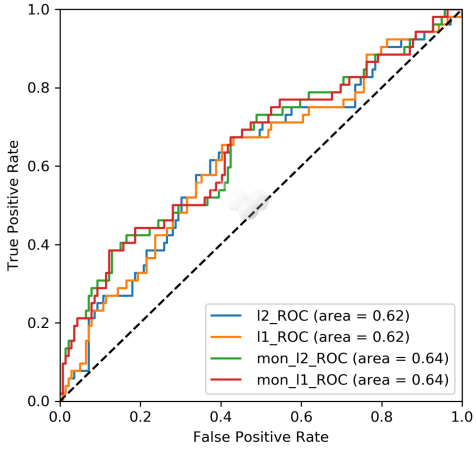
Задача ЭА–ЖК



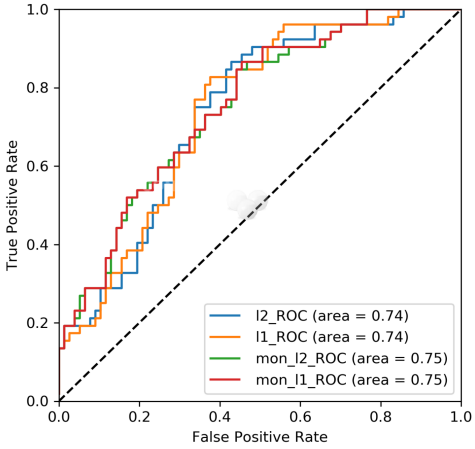
Задача ЭА–ИБ



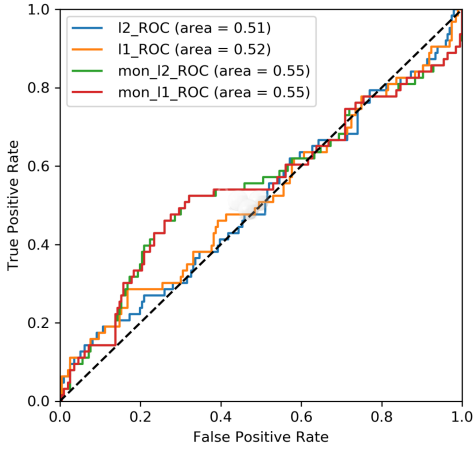
Задача ЭА–МК



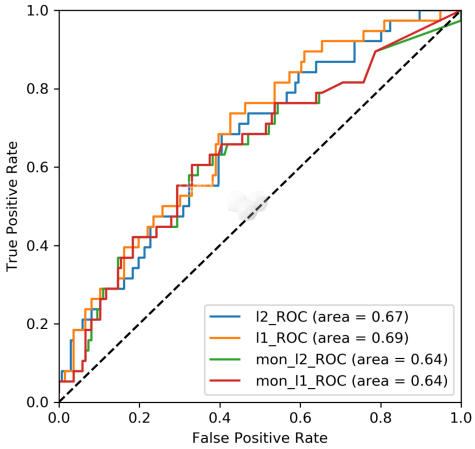
Задача ЭА–ММ



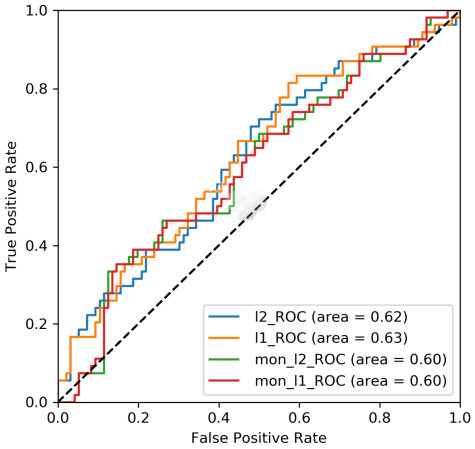
Задача ЭА–СД



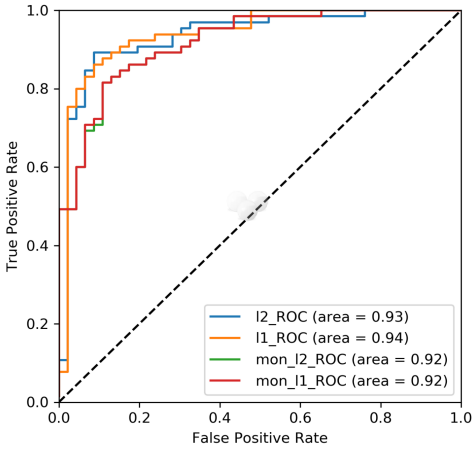
Задача ЭА–УЩ



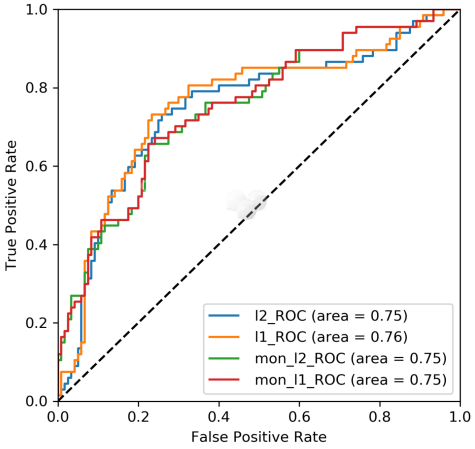
Задача ЭА–ХГ



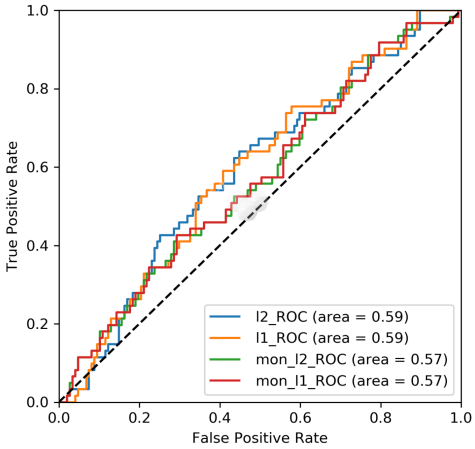
Задача ЭА–ХХ



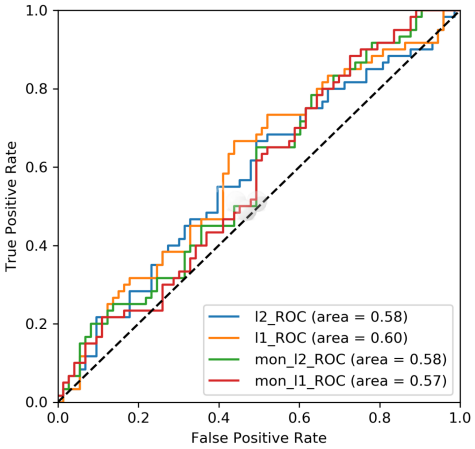
Задача АП–АЗ



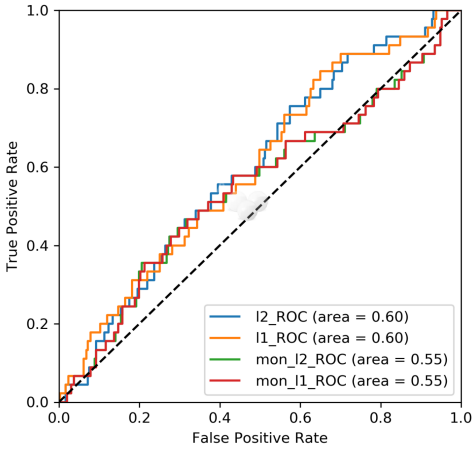
Задача АП–ВД



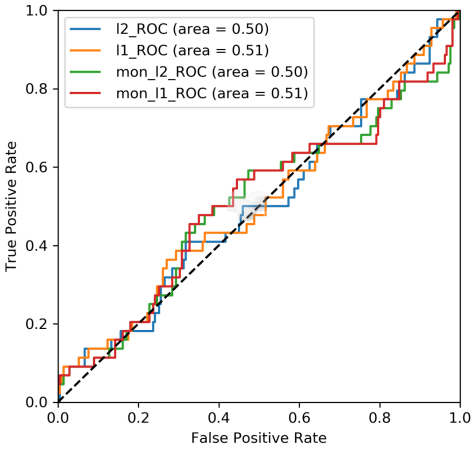
Задача АП–ГБ



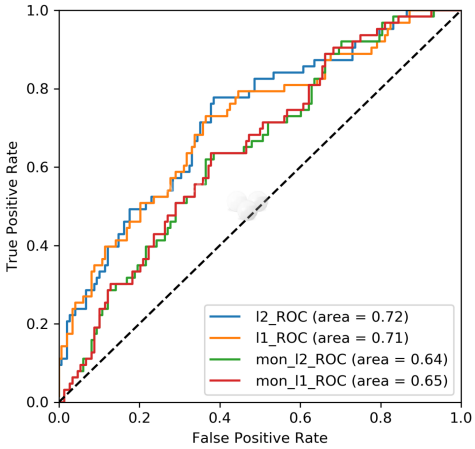
Задача АП–ЖК



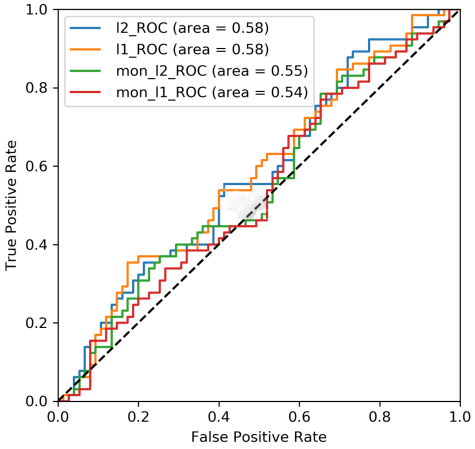
Задача АП–ИБ



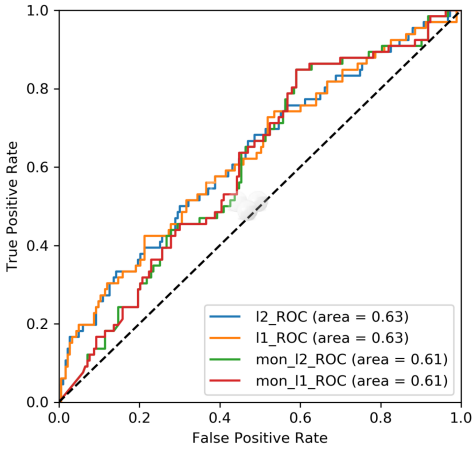
Задача АП–МК



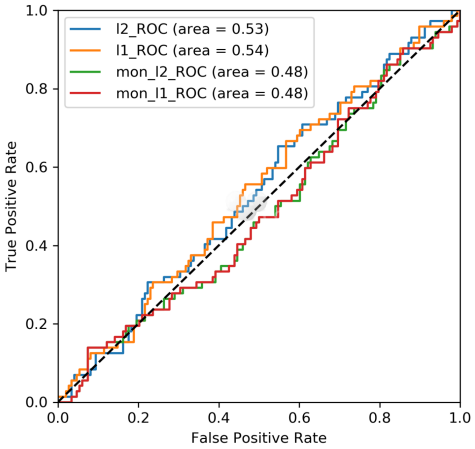
Задача АП–ММ



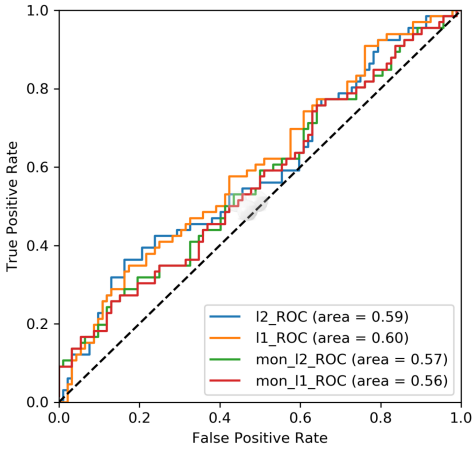
Задача АП–СД



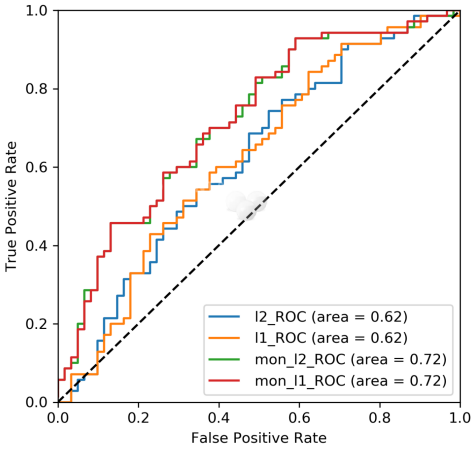
Задача АП–УЩ



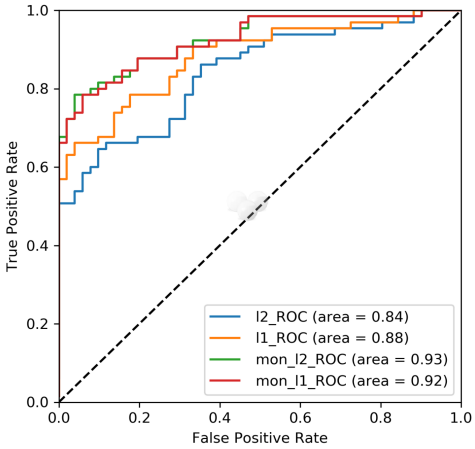
Задача АП–ХГ



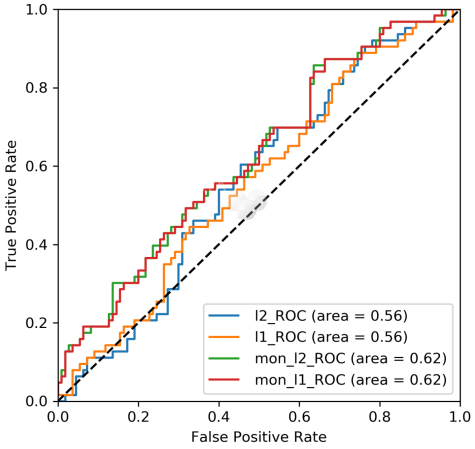
Задача АП–ХХ



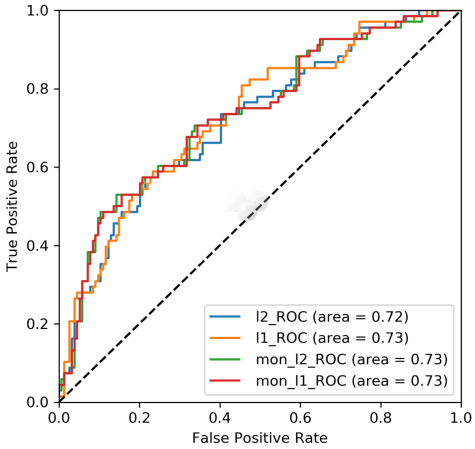
Задача АП–ЭА



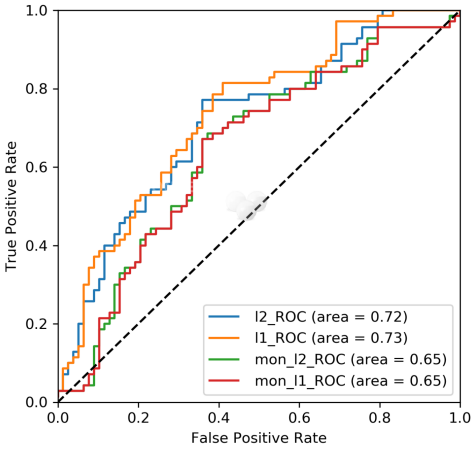
Задача АХ–АЗ



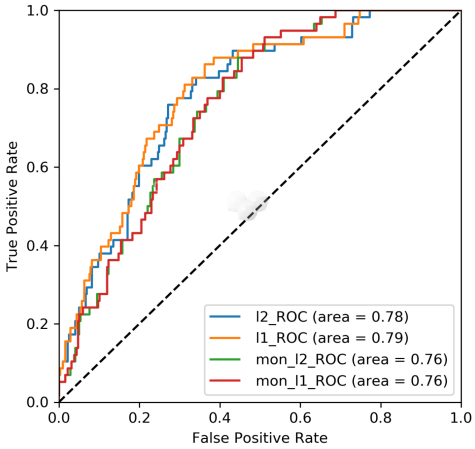
Задача АХ–ВД



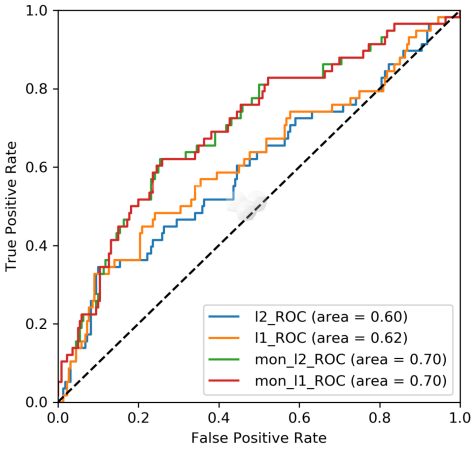
Задача АХ–ГБ



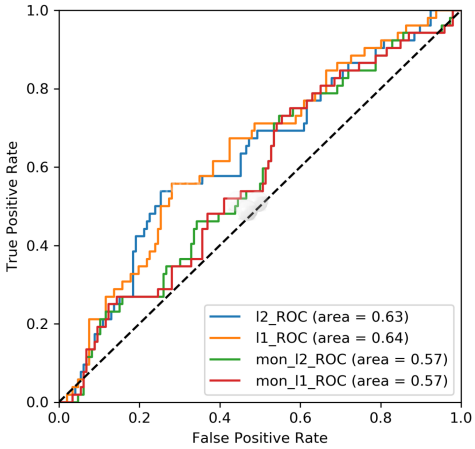
Задача АХ–ЖК



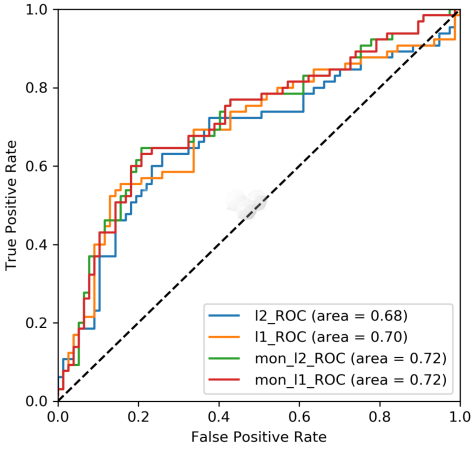
Задача АХ–ИБ



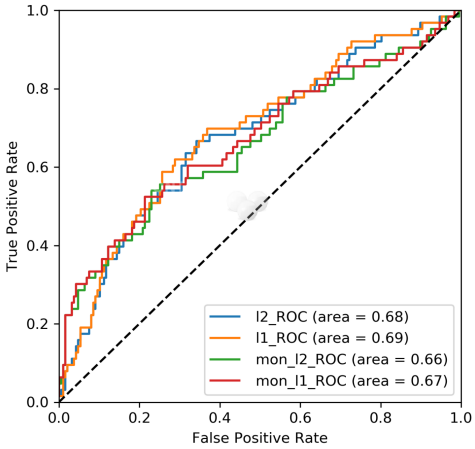
Задача АХ–МК



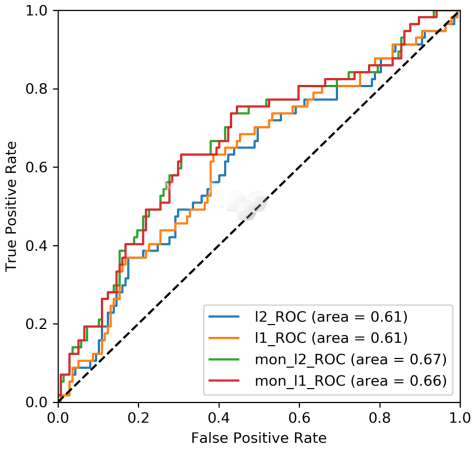
Задача АХ–ММ



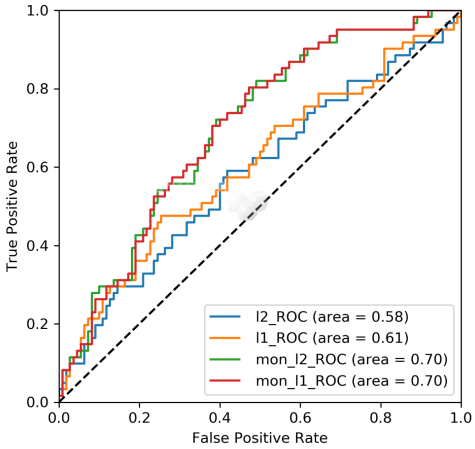
Задача АХ–СД



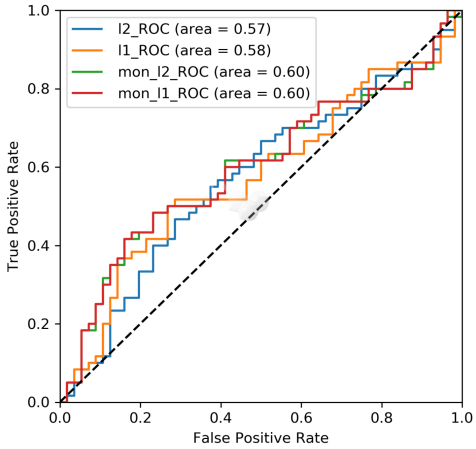
Задача АХ–УЩ



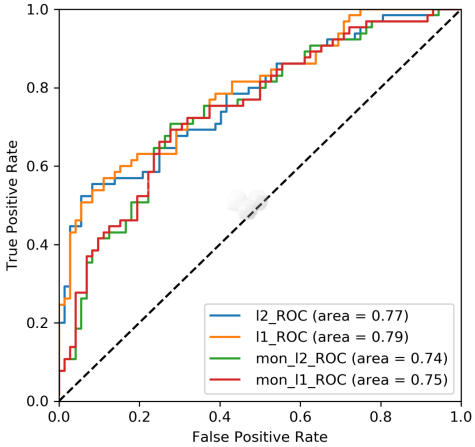
Задача АХ–ХГ



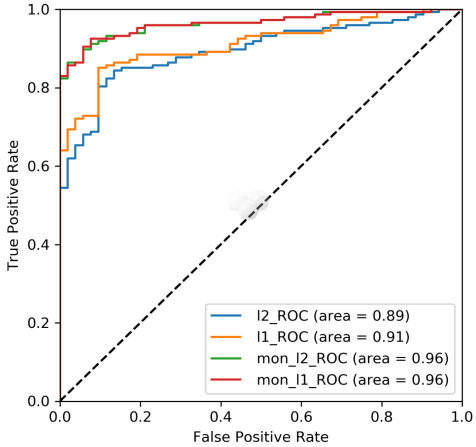
Задача АХ–ХХ



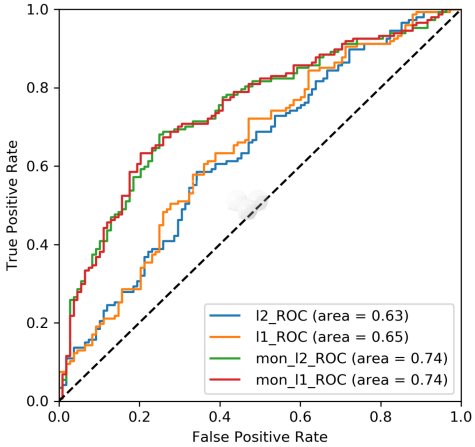
Задача АХ–ЭА



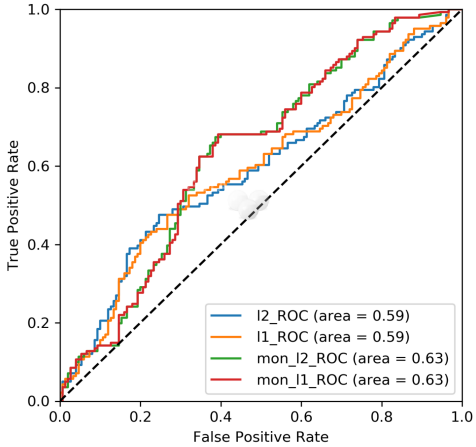
Задача АХ–АП



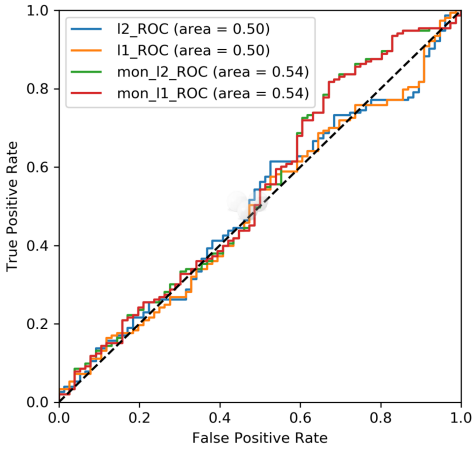
Задача ЯБ–АЗ



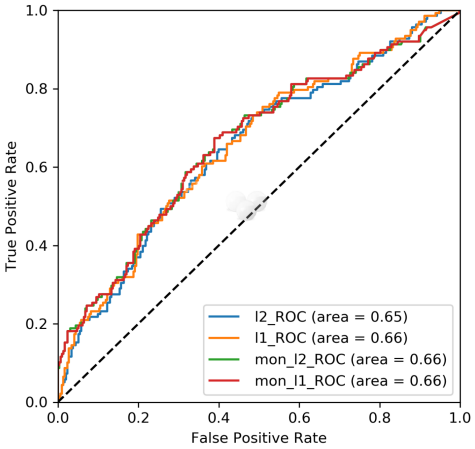
Задача ЯБ–ВД



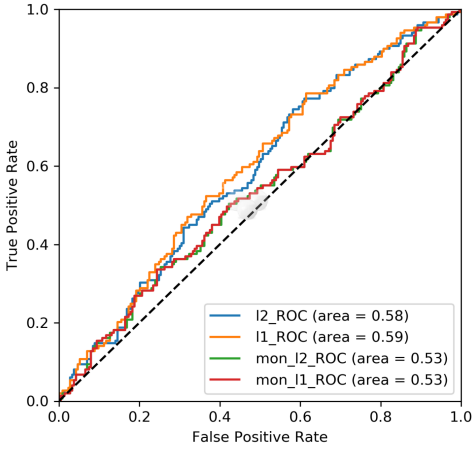
Задача ЯБ–ГБ



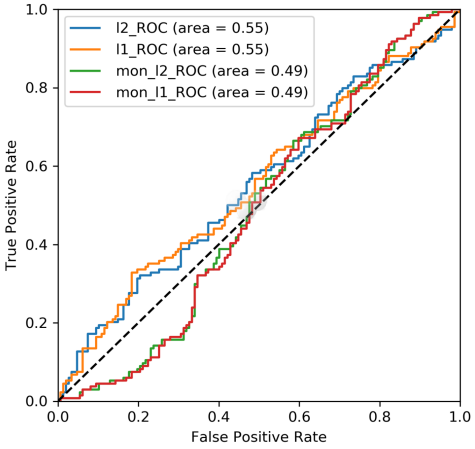
Задача ЯБ–ЖК



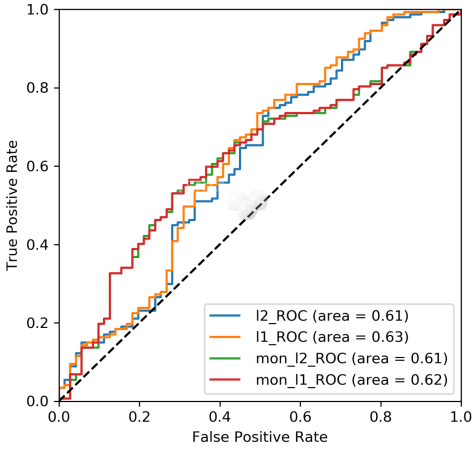
Задача ЯБ–ИБ



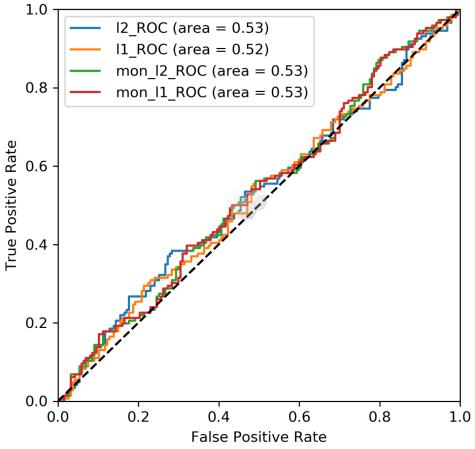
Задача ЯБ–МК



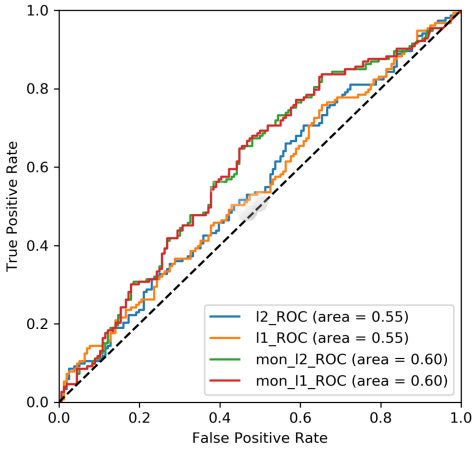
Задача ЯБ–ММ



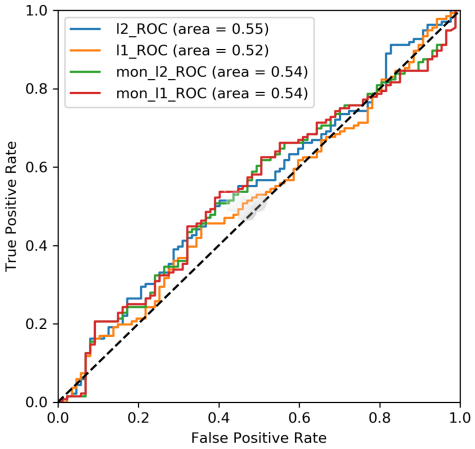
Задача ЯБ–СД



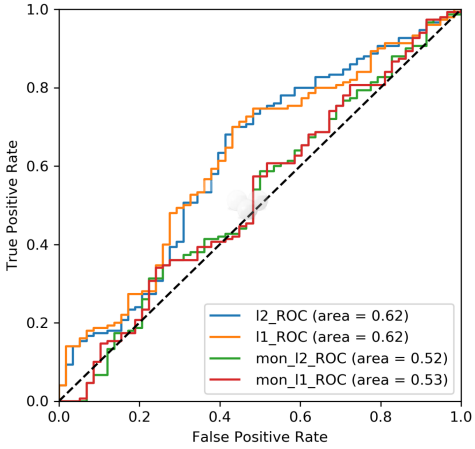
Задача ЯБ–УЩ



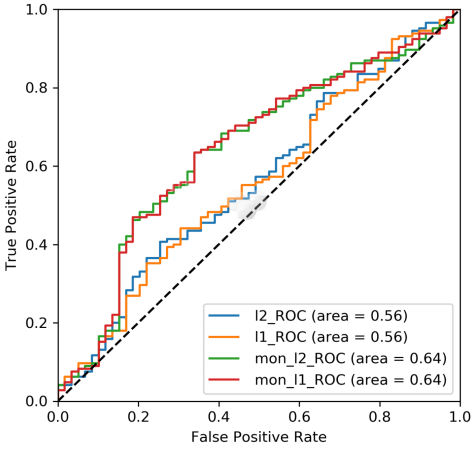
Задача ЯБ–ХГ



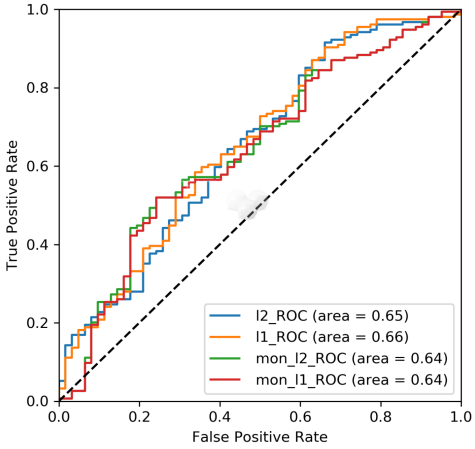
Задача ЯБ–ХХ



Задача ЯБ–ЭА



Задача ЯБ–АП



Задача ЯБ–АХ