Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТКИ 974 ГРУППЫ

«Построение интегральных индикаторов по частично упорядоченным множествам экспертных оценок»

Выполнила:
студентка 4 курса 974 группы
Медведникова Мария Михайловна

Научный руководитель:
к.ф-м.н., н.с. ВЦ РАН
Стрижов Вадим Викторович

Москва, 2013

Содержание

Аннотация

Рассмотрена задача многоклассовой монотонной классификации в ранговых шкалах. Используется нахождение Парето-оптимального фронта. Предложен метод агрегирования признаков с учетом их иерархии и важности, определенной экспертным путем. Рассмотрен метод проверки выполнения отношения транзитивности для объектов разных классов. Работа проиллюстрирована задачей прогнозирования статуса редких видов, включенных в Красную книгу РФ.

Ключевые слова: *многоклассовая классификация, ранговые шкалы, Парето-оптимальный фронт, агрегирование признаков.*

Введение

Актуальность темы. На сегодняшний день задача построения интегральных индикаторов встречается в различных прикладных областях. С их помощью оценивают различные экономические показатели, эффективность научной работы, качество научных изданий, строят рейтинги организаций.

Как правило данными, на основе которых строится индикатор, являются экспертные оценки различных признаков группы объектов. Обычно признаки, значения которых определялись экспертами, описываются в ранговых шкалах. Часто эксперт может указать признаки, которые считает важнее других, либо указать иерархию признаков. Если количество признаков, описывающих объекты велико по сравнению с числом объектов, такая информация может быть использована для отбора признаков или уменьшения их числа путем различных сверток.

Цель работы. Построить алгоритм многоклассовой монотонной классификации объектов, описанных в ранговых шкалах, позволяющий учитывать экспертную информацию об иерархии признаков и их важности.

Методы исследований. При построении алгоритма использовались элементы теории бинарных отношений, методы обработки нечисловой информации, методы решения задач классификации с учетом предпочтений, алгоритм Парето-расслоения. Для программной реализации разработанного алгоритма использовалась среда MATLAB.

Научная новизна.

- Предложен метод свертки наборов признаков, указанных экспертом с целью уменьшения количества признаков и ликвидации признаков, описывающих логически связанные между собой свойства объектов;

- разработан алгоритм монотонной классификации с использованием Парето-оптимальных фронтов двух типов, учитывающий экспертную информацию о важности признаков;

- разработан двухуровневый алгоритм классификации, учитывающий экспертную иерархию признаков.

Практическая ценность. Разработан программный модуль, который

- строит модель интегрального индикатора;
- по экспертным оценкам прогнозирует значение интегрального индикатора;
- визуализирует результаты.

Положения, выносимые на защиту:

- Алгоритм монотонной многоклассовой классификации с использованием Парето-оптимальных фронтов двух типов, учитывающий экспертную информацию о важности признаков;
- двухуровневый алгоритм классификации, учитывающий экспертную иерархию признаков.

Апробация. Результаты квалификационной работы бакалавра были использованы для решения задачи определения статуса редких видов, включенных в Красную книгу РФ [1,2] при участии экспертов министерства природных ресурсов и экологии и IUCN.

Обзор литературы. Будем называть объектом применительно к данной задаче вид, включенный в Красную книгу, а признаком — некоторый параметр, характеризующий вид (например, численность, площадь ареала, и т. д.) и принимающий значения из множества, на котором введено отношение частичного порядка. Прикладная задача заключается в определении категории риска для редкого вида по его признаковому описанию.

Рассматриваемая задача построения интегрального индикатора является задачей монотонной классификации (задачей ранжирования [7,8]). Задачи подобного типа часто возникают в сфере информационного поиска. Для их решения используют

ранговую регрессию [9], модифицированный алгоритм SVM [10] и модифицированный бустинг [11]. Они отличаются от классических алгоритмов функциями потерь, учитывающими монотонность классификации.

Описание вида составляется экспертом по утвержденному набору признаков. Количество признаков сопоставимо с количеством объектов, доступных для обучения алгоритма. Задачи с избыточным числом признаков рассматриваются в работах [3–6]. Предлагаются различные подходы к решению проблемы чрезмерно большого числа признаков: различные методы отбора признаков [3], метод главных компонент [4], ассоциативные правила, учитывающие информативность признаков [5], свертки нескольких признаков в один [6].

Признаки, использованные для описания объектов принимают значения из множеств, на которых задано отношение порядка. Отношение частичного порядка является одним из видов бинарных отношений, свойства которых рассматриваются в [12]. Объекты, описанные в ранговых шкалах, не являются точками в некотором линейном пространстве, они представляют собой объекты нечисловой природы. Подходы к обработке нечисловой информации описаны в [13, 14]. В рассматриваемой прикладной задаче имеется экспертная информация о важности признаков относительно друг друга, то есть над множеством признаков тоже задано бинарное отношение предпочтения. Также над признаками задана иерархия и имеется информация об относительной важности поднаборов признаков (задано отношение предпочтения над множеством поднаборов признаков). Задача монотонной классификации объектов нечисловой природы с учетом предпочтений освещается в [15–25]. Основу методов ее решения составляют попарные сравнения [22], также используется бустинг [11].

Для построения интегрального индикатора в рассматриваемой задаче используется алгоритм многоклассовой монотонной Парето-классификации [26]. Алгоритм подробно описан в статье [31]. Логическое обоснованние принципа Парето представлено работе [27]. Предполагается, что для более устойчивой работы алгоритма целесообразно минимизировать количество объектов, входящих в Парето-фронт. Подходы к сужению множества Парето описаны в [28, 29]. В данной работе предлагается сузить множество Парето путем учитывания экспертной информации о важности признаков [30] с помощью модификации отношения доминирования объектов [26].

Проблема избыточного числа признаков решается путем комбинирования двух приемов. В первую очередь проводится свертка групп признаков [6] с целью получения меньшего числа признаков, описанных в более мощных шкалах, затем учитывается иерархия признаков. Для этого строится ряд классификаторов, обученных на различных непересекающихся поднаборах признаков. Результаты их работы интерпретируются как новые признаковые описания объектов, которые передаются на вход «внешнему» классификатору, учитывающему важность использованных на предыдущем шаге поднаборов признаков.

Построенный алгоритм сравнивается с алгоритмом решающего дерева, алгоритмом криволинейной регрессии [33], с алгоритмома на основе копул [34] и конусов [35].

1 Постановка задачи, исходные гипотезы, входные данные

1.1 Целеполагание и исходные гипотезы

Согласно законодательству [32], Красную книгу необходимо пересматривать не реже одного раза в десять лет с целью ревизии статуса видов. В настоящее время Красная книга включает более четырехсот видов животных. Каждому виду присваивается один из шести статусов: 0 — вероятно исчезнувшие; 1 — находящиеся под угрозой исчезновения; 2 — сокращающиеся в численности; 3 — редкие; 4 — неопределенные по статусу; 5 — восстанавливаемые и восстанавливающиеся.

Предполагается, что текущая версия Красной книги РФ составлена «в целом» непротиворечиво, то есть, существует соответствие между описанием вида и его статусом.

В ходе работы будут использованы следующие предположения о составе и свойствах признаков:

- оцениваемый экспертами состав признаков считается исчерпывающим для получения модели;
- учитывается структура на признаках;

- на значениях признаков задано отношение полного порядка (кроме пропущенного значения);
- выполняется правило «the bigger the better», то есть большему (благоприятному) значению признака соответствует больший (благоприятный) статус вида;
- различные экспертные оценки по одному и тому же виду приветствуются;
- каждый из признаков принимает на выборке все допустимые значения и только их, при нарушении условия ставится вопрос о корректности анкеты;
- признак отвергается, если более половины значений пропущены.

Также будет изменен список статусов. Не будет рассматриваться категория 0 ввиду иного способа категоризации. Статусы 4, 5 требуется исключить, перераспределив виды, входящие в них, по другим статусам либо исключив их из Красной книги.

Требования, предъявляемые к модели.

- Корректное использование шкал экспертных оценок;
- оптимальная сложность;
- достаточно хорошее описание текущих категорий видов (с учетом изменений в шкале категорий).

1.2 Постановка задачи

Дано множество пар $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$, состоящее из объектов $\mathbf{x} = [\chi_1, \ldots, \chi_j, \ldots, \chi_d]^\intercal$, $j \in \mathcal{J} = \{1, \ldots, d\}$, описанных в ранговых шкалах, $\chi_j \in \mathbb{L}_j = \{l_k : l_1 \prec \cdots \prec l_{k_j}\}$. Требуется построить монотонную функцию

$$\varphi \colon \mathbf{x} \mapsto \hat{y}, \tag{1}$$

определенную на всем множестве $\mathbb{X} = \mathbb{L}_1 \times \cdots \times \mathbb{L}_d$, $\mathbb{X} \ni \mathbf{x}$, и принимающую значения из множества $\mathbb{Z} = \{1, \ldots, z\}$, $1 \prec \ldots \prec z$. Эта функция должна доставлять минимум

функции ошибки

$$S(\varphi) = \frac{1}{m} \sum_{i \in \mathcal{I}} r(y_i, \hat{y}_i), \tag{2}$$

где $\hat{y}_i = \varphi(\mathbf{x}_i)$; r задает расстояние между метками упорядоченного множества.

Таблица 1: Матрица отношения порядка

Метки	1	2	...	$z-1$	z
1	0	0	...	0	0
2	1	0	...	0	0
...			...		
$z-1$	1	1	...	0	0
z	1	1	...	1	0

Если изобразить отношение порядка на множестве с помощью бинарной матрицы, то эта матрица будет нижне-треугольной. Поставим в соответствие метке y_i бинарную строку str_i из табл. 1. Тогда расстояние r между метками y_i и y_j будет задаваться расстоянием Хэмминга между бинарными векторами

$$r(y_i, y_j) = R_{\text{Ham}}(str_i, str_j). \tag{3}$$

1.3 Структура данных

Эксперимент проводится на выборке, состоящей из экспертных оценок видов Красной книги РФ и экспертных оценок важности признаков. В выборку вошли 102 объекта из трех категорий риска, описанные 102-мя признаками. Признаки представлены в таблице на рис. 1. Все признаки, входящие в анкету, заполняемую экспертом, делятся на пять поднаборов, которые распределены по разным столбцам таблицы. В каждую ячейку таблицы входят от одного до четырех признаков, описывающих одно свойство вида.

Выборка представляет собой множество пар $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$, состоящее из объектов $\mathbf{x} = [\chi_1, \ldots, \chi_j, \ldots, \chi_d]^\intercal$, $j \in \mathcal{J} = \{1, \ldots, d\}$, описанных в ранговых шкалах, $\chi_j \in \mathbb{L}_j = \{l_k : l_1 \prec \cdots \prec l_{k_j}\}$. Каждый признак χ_j принимает значение из множества $\mathbb{L}_j$, на элементах которого, кроме пропущенного значения, задано отношение линейного порядка. Для наглядности будем считать, что элементы из

Биологическое состояние	Суммарные угрозы	Значимость	Защищенность	Готовность
Численность	Климатические изменения	Возможные потери	Конвенции, договора, законодательство	Искусственное воспроизведение популяций
Тенденция изменения численности	Геологические катастрофы	Для ареала в России	применение, достаточность, эффективность	Реинтродукция в природу
Темп изменения численности	Биоценотические факторы	Биоценотическая роль	Стратегии, менеджмент-планы	Сохранение EX-SITU
Тенденция изменения темпа численности	Разрушение местообитаний	Ресурсная значимость	применение, достаточность, эффективность	Необходимые затраты на обеспечение сохранения/восстановления
Встречаемость/Плотность	Добыча (сбор)	Научная значимость	ООПТ	Степень изученности
Тенденция изменения плотности	Загрязнения	Индикаторная значимость	применение, достаточность, эффективность	Налаженность мониторинга
Общая площадь ареала	Интродукция чужеродных видов		Территории международного значения, Ключевые территории	
Тенденция изменения площади ареала	Случайная гибель		применение, достаточность, эффективность	
Структура ареала	Беспокойство		Восстановление и сохранение местообитаний	
Тенденция изменения структуры ареала	Эксплуатация кормовой базы		применение, достаточность, эффективность	
Популяционная структура вида			Искусственное воспроизводство популяций	
Тенденция изменения популяционной структуры			применение, достаточность, эффективность	
Генетическое разнообразие			Реинтродукция (реакклиматизация)	
Тенденция изменения генетического разнообразия			применение, достаточность, эффективность	
Половозрастная и социальная структура			Создание новых популяций	
Тенденция изменения половозрастной структуры			применение, достаточность, эффективность	
Физиологическое состояние			Регуляция использования и торговли	
Тенеденция изменения физиологического состояния			применение, достаточность, эффективность	
Состояние местообитаний			Управление воспроизводством	
Тенденция изменения состояния местообитания			применение, достаточность, эффективность	
			Борьба с болезнями и паразитами	
			применение, достаточность, эффективность	
			Регуляция численности	
			применение, достаточность, эффективность	
			Сохранение генетических материалов	
			применение, достаточность, эффективность	
			Содержание и разведение в неволе	
			применение, достаточность, эффективность	
			Введение в культуру	
			применение, достаточность, эффективность	
			Экологическое образование	
			применение, достаточность, эффективность	
			Эколого-просветительская работа	
			применение, достаточность, эффективность	
			Capacity Building/Training	
			применение, достаточность, эффективность	
			Экологическая пропаганда	
			применение, достаточность, эффективность	
			Эколого-художественная деятельность	
			применение, достаточность, эффективность	

Рис. 1: Список признаков — полей анкеты для описания вида экспертом

$\mathbb{L}_j$ тождественны элементам подмножества натуральных чисел: $l_1 = 1, \ldots, l_{k_j} = k_j$. Пропущенное значение имеет метку «0». На множестве $\mathbb{Y} = \{1, 2, 3\}$ меток классов y также задано отношение порядка: $1 \prec 2 \prec 3$.

1.4 Предобработка данных

В признаковых описаниях объектов выборки присутсвуют пропуски. Были удалены признаки «Генетическое разнообразие», «Тенденция изменения генетического разнообразия» и «Сохранение Ex-situ», так как их значения пропущены более, чем у 50% объектов выборки.

2 Иерархия признаков

На признаках введена многоуровневая иерархия, отраженная в схеме на рис. 2. На схеме показано, как объединяются признаки по типам свойств, которые они описавыют. В работе использована не вся структура иерархии, только элементы, заключенные в эллипсы. Этим поднаборам признаков соответсвуют столбцы таблицы на рис. 1. Внутри каждого поднабора иерархия не учитывается, то есть признаки внутри поднабора не делятся на какие-либо меньшие поднаборы.

2.1 Поднаборы признаков

Поднаборы признаков представляют собой элементы верхнего уровня используемой иерархии и соответствуют столбцам таблицы на рис. 1. Для разделения набора всех признаков на поднаборы множество индексов $\mathcal{J}$ разбивается на пять непересекающихся подмножеств $\mathcal{J} = \mathcal{A}_1 \sqcup \ldots \sqcup \mathcal{A}_5$, каждое из которых включает индексы признаков, описывающих определенную группу свойств объектов. Имеется экспертная информация о важности поднаборов признаков относительно друг друга. Будем обозначать $A_i \succ A_j$, если поднабор A_i важнее поднабора A_j. Важность поднаборов признаков $\mathcal{A}_1 \succ \ldots \succ \mathcal{A}_5$ убывает по столбцам слева направо.

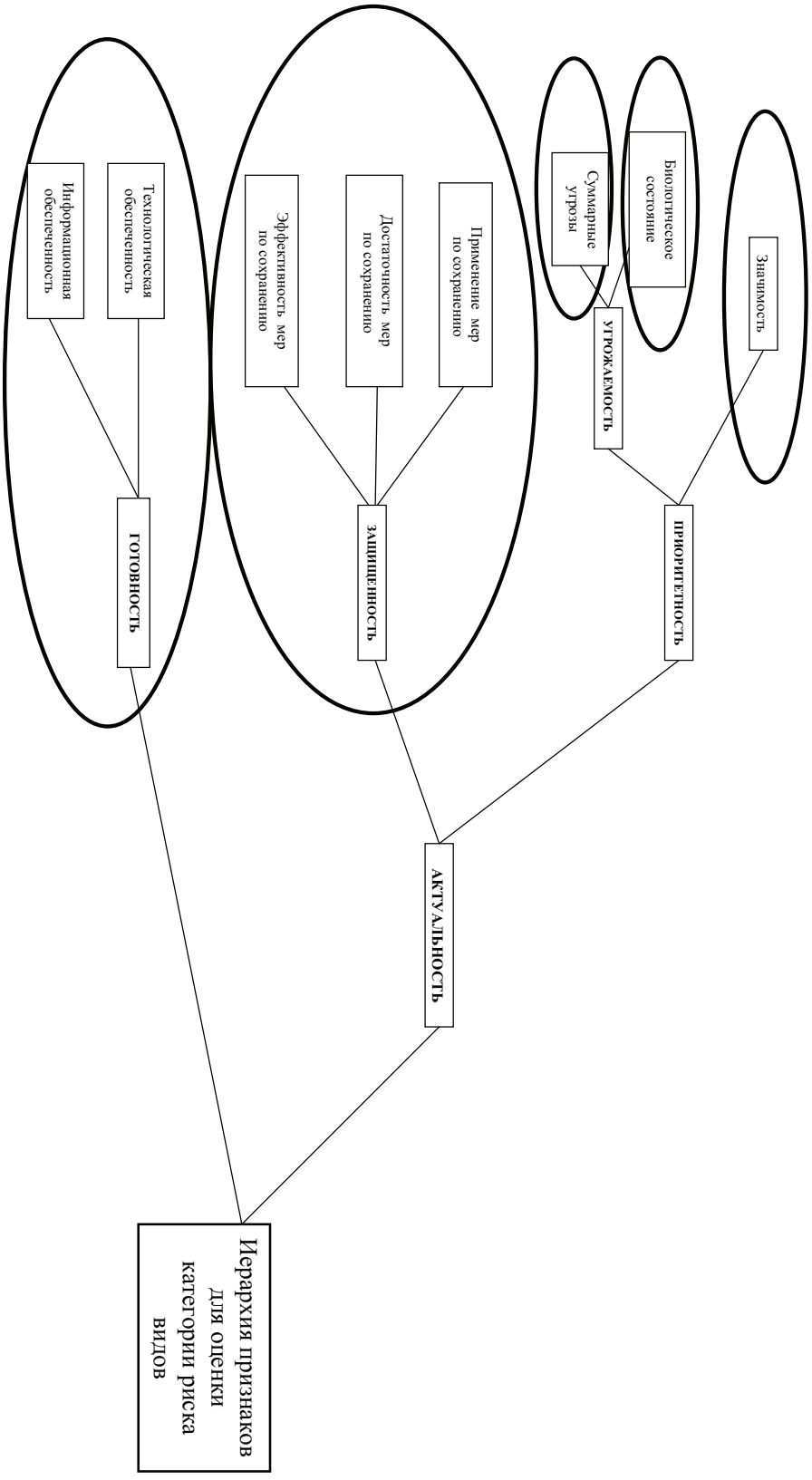


Рис. 2: Структура, введенная экспертом на множестве признаков

2.2 Свертки признаков

Свертки признаков представляют собой элементы нижнего уровня иерархии и соответствуют ячейкам таблицы на рис. 1.

Пусть $\{\chi_1, \ldots, \chi_p\}$ — признаки, входящие в один набор для агрегирования (находящиеся в одной ячейке таблицы на рис. 1) и описанные в ранговых шкалах $\chi_j \in \mathbb{L}_j = \{1, \ldots, k_j\}, \quad j = 1, \ldots, p$. Значение агрегированного признака ψ находится как сумма значений признаков χ_j, $j = 1, \ldots, p$, нормированная таким образом, что она принимает значения от 1 до $\sum_{j=1}^{p} k_j - p + 1$:

$$\psi = \sum_{j=1}^{p} \chi_j - p + 1.$$

После проведения процедур предобработки данных и агрегирования признаков исходное множество индексов признаков $\mathcal{J} = \{1, \ldots, d\}$ преобразуется в множество $\mathcal{J}' = \{1, \ldots, d'\}$ с меньшим числом элементов. Вводится частичный порядок на каждом подмножестве индексов, соответствующем поднабору признаков $j \in \mathcal{J}_{\mathcal{A}_i} \subset \mathcal{J}'$, $i = 1, \ldots, 5$. Будем обозначать $j_r \succ_j j_t$, если признак j_r важнее признака j_t, то есть в предоставленной экспертами информации содержится такое утверждение. На рис. 3 изображены таблицы, отражающие важность признаков в поднаборах «Биологическое состояние», «Суммарные угрозы» и «Защищенность». В ячейке таблицы стоит 1, если признак в строке важнее, чем признак в столбце, и 0 в противном случае. Представленные таблицы для данных поднаборов были заполнены экспертом. Информация о поднаборах «Значимость» и «Готовность» не была предоставлена. Соответствующие им таблицы не приведены, поскольку признаки, входящие в эти поднаборы, считаются несравнимыми и таблицы заполнены нулями.

Считается, что в предоставленной информации о важности признаков нет противоречий, если введенное отношение частичного порядка не объединяет признаки в циклы $j_{i_1} \succ_j \ldots \succ_j j_{i_n} \succ j_{i_1}$. Если какие-либо признаки объединяются в цикл, то отношение порядка между ними не учитывается и признаки считаются несравнимыми по важности. На рис. 3(b) желтым цветом покрашены ячейки, соответсвующие признакам, информация о важности которых противоречива.

Если признак в строке важнее, чем в столбце, нужно поставить 1	Численность	Темп изменения численности	Встречаемость/ Плотность	Общая площадь ареала	Структура ареала	Популяционная структура вида	Генетическое разнообразие	Половозрастная структура	Физиологическое состояние	Состояние местообитаний
Численность	0	1	1	1	1	1	1	1	1	1
Темп изменения численности	0	0	1	1	1	1	1	1	1	1
Встречаемость/ Плотность	0	0	0	1	1	1	1	1	1	1
Общая площадь ареала	0	0	0	0	1	1	1	1	1	1
Структура ареала	0	0	0	0	0	1	1	1	1	0
Популяционная структура вида	0	0	0	0	0	0	1	1	1	0
Генетическое разнообразие	0	0	0	0	0	0	0	0	0	0
Половозрастная и социальная	0	0	0	0	0	0	1	0	1	0
Физиологическое состояние	0	0	0	0	0	0	1	0	0	0
Состояние местообитаний	0	0	0	0	1	1	1	1	1	0

(a) Биологическое состояние

Если признак в строке важнее, чем в столбце, нужно поставить 1	Климатические изменения	Геологические катастрофы	Биоценотические факторы	Разрушение местообитаний	Добыча (сбор)	Загрязнения	Интродукция чужеродных видов	Случайная гибель	Беспокойство	Эксплуатация кормовой базы
Климатические изменения	0	1	0	0	0	0	1	1	1	0
Геологические катастрофы	0	0	0	0	0	0	0	1	0	0
Биоценотические факторы	1	1	0	0	0	0	1	1	1	0
Разрушение местообитаний	1	1	1	0	0	1	1	1	1	0
Добыча (сбор)	1	1	1	0	0	1	1	1	1	0
Загрязнения	1	1	1	0	0	0	1	1	1	0
Интродукция чужеродных видов	0	0	0	0	0	0	0	1	1	0
Случайная гибель	0	0	0	0	0	0	0	0	0	0
Беспокойство	0	1	0	0	0	0	1	1	0	0
Эксплуатация кормовой базы	1	1	1	0	0	1	1	1	1	0

(b) Суммарные угрозы

Если признак в строке важнее, чем в столбце, нужно поставить 1	Конвенции, договора, законодательство	Стратегии, менеджмент-планы	ООПТ	территории	Восстановление и сохранение местообитаний	Искусственное воспроизводство популяций	Реинтродукция (реакклиматизация)	Создание новых популяций	Регуляция использования и торговли	Управление воспроизводством	Борьба с болезнями и паразитами	Регуляция численности	Сохранение генетических материалов	Содержание и разведение в неволе	Введение в культуру	Экологическое образование	Эколого-просветительская работа	Capacity Building/Training	Экологическая пропаганда	Эколого-художественная деятельность
Конвенции, договора, законодательство	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Стратегии, менеджмент-планы	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ООПТ	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Территории международного значения, Ключевые территории	0	0	0	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
Восстановление и сохранение местообитаний	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Искусственное воспроизводство популяций	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0
Реинтродукция (реакклиматизация)	0	0	0	0	0	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0
Создание новых популяций	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
Регуляция использования и торговли	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
Управление воспроизводством	0	0	0	0	0	1	0	1	0	0	1	1	0	0	1	0	0	0	0	0
Борьба с болезнями и паразитами	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Регуляция численности	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Сохранение генетических материалов	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
Содержание и разведение в неволе	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0	0	0	0	0
Введение в культуру	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Экологическое образование	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	1
Эколого-просветительская работа	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	0	0	0	1
Capacity Building/Training	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	1	0	0	0	1
Экологическая пропаганда	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	1
Эколого-художественная деятельность	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	0

(c) Защищенность

Рис. 3: Информация о важности признаков

2.3 Построение интегрального индикатора с учетом иерархии признаков

В ходе работы строится монотонный классификатор (1) для каждого подмножества индексов признаков $\mathcal{A}_1, \ldots, \mathcal{A}_5$ из множества $\mathcal{J}'$. Таким образом для каждого объекта $\mathbf{x}$ получим набор промежуточных результатов классификации на поднаборах признаков $y_i = \varphi_{\mathcal{A}_i}(\mathbf{x}), i = 1, \ldots, 5$. Будем считать вектор $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ новым признаковым описанием объекта $\mathbf{x}$:

$$\mathbf{x} \mapsto \mathbf{y} = \begin{pmatrix} \varphi_{\mathcal{A}_1}(\mathbf{x}) \\ \ldots \\ \varphi_{\mathcal{A}_5}(\mathbf{x}) \end{pmatrix}.$$

Для окончательного принятия решения построим еще один монотонный классификатор (1), который принимает на вход вектор промежуточных результатов классификации $\mathbf{y}$ и информацию о важности новых признаков $\mathcal{A}_1 \succ \ldots \succ \mathcal{A}_5$, содержащихся в $\mathbf{y}$:

$$s = \varphi(\mathbf{y}),$$

где s — окончательный результат классификации.

3 Парето-классификация для случая двух классов

Для решения задачи двухклассовой классификации предлагается построить монотонную функцию $f\colon \mathbf{x} \mapsto \hat{y}$, определенную на всем множестве

$$\mathbb{X}' = \mathbb{L}_1 \times \cdots \times \mathbb{L}_{d'}, \tag{4}$$

$\mathbb{X}' \ni \mathbf{x}$, и принимающую значения из множества $\mathbb{Y} = \{0, 1\}$, $0 \prec 1$. Эта функция должна доставлять минимум функции ошибки

$$S(f) = \frac{1}{m} \sum_{i \in \mathcal{I}} r(y_i, \hat{y}_i),$$

где $\hat{y}_i = f(\mathbf{x}_i)$, r задано в (3).

Решим задачу нахождения функции $f(\mathbf{x})$ с помощью разделимой выборки $\hat{\mathfrak{D}} = \{(\mathbf{x}_i, y_i)\}$, $i \in \hat{\mathcal{I}} \subseteq \mathcal{I}$: предполагается, что каждому из классов y соответствует выпуклая оболочка POF, заданная отношением доминирования «$\succ$» элементов, и эти

оболочки не пересекаются. Искомая функция f будет сначала определена (5) на множестве объектов $\{\mathbf{x}_i : i \in \hat{\mathcal{I}}\}$ с индексами $i \in \hat{\mathcal{I}} \subseteq \mathcal{I}$, а затем доопределена (4) на всем множестве $\mathbb{X}$.

3.1 Отношение доминирования без учета важности признаков

Введем на объектах каждого из классов отношение доминирования. Разобьем множество индексов $\hat{\mathcal{I}}$ объектов разделимой выборки $\hat{\mathfrak{D}}$ на два подмножества $\hat{\mathcal{I}} = \mathcal{N} \bigsqcup \mathcal{P}$ так, что $y_n = 0$, а $y_p = 1$, $n \in \mathcal{N}, p \in \mathcal{P}$. Введем на множествах $\{\mathbf{x}_n : n \in \mathcal{N}\}$ и $\{\mathbf{x}_p : p \in \mathcal{P}\}$ отношения доминирования $\succ_n$ и $\succ_p$. Объект $\mathbf{x}_n$ n-доминирует объект $\mathbf{x}_i$, если справедливы неравенства:

$$\mathbf{x}_n \succ_n \mathbf{x}_i, \quad \text{если} \quad x_{nj} \geq x_{ij} \quad \text{для всех} \quad j \in \mathcal{J}'.$$

Аналогично, объект $\mathbf{x}_p$ p-доминирует $\mathbf{x}_k$, если справедливы неравенства:

$$\mathbf{x}_p \succ_p \mathbf{x}_k, \quad \text{если} \quad x_{pj} \leq x_{kj} \quad \text{для всех} \quad j \in \mathcal{J}'.$$

Будем считать, что объект не доминирует сам себя ни в одном из смыслов:

$$\mathbf{x} \not\succ_n \mathbf{x}, \quad \mathbf{x} \not\succ_p \mathbf{x}.$$

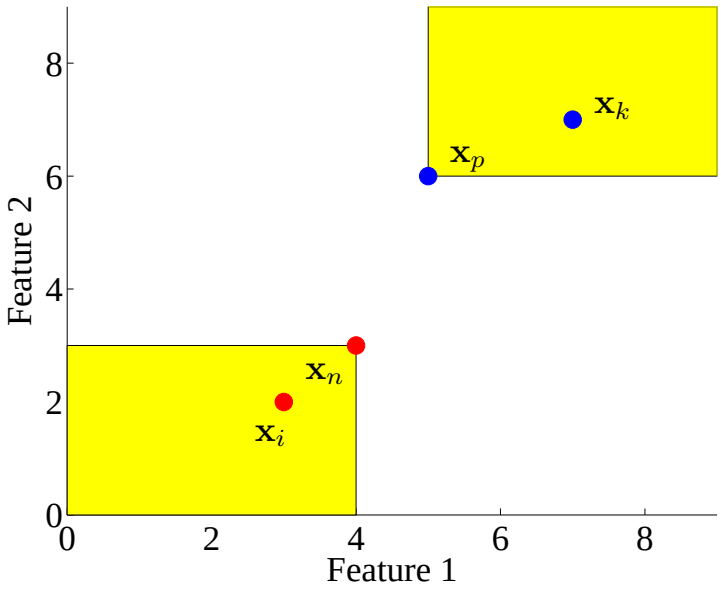


Рис. 4: Доминирование без учета важности признаков

На рис. 4 приведен пример доминирования для случая двух признаков. По осям отложены значения признаков, цветом показаны область n-доминирования объекта со значениями признаков (4;3) и область p-доминирования объекта со значениями признаков (5;6). Объекты, попадающие в области, закрашенные желтым цветом, доминируются в соответствующем смысле рассмотренными объектами.

3.2 Отношение доминирования с учетом важности признаков

Введем на объектах каждого из классов отношение обобщенного доминирования. Разобьем множество индексов $\hat{\mathcal{I}}$ объектов разделимой выборки $\hat{\mathfrak{D}}$ на два подмножества $\hat{\mathcal{I}} = \mathcal{N} \bigsqcup \mathcal{P}$ так, что $y_n = 0$, а $y_p = 1$, $n \in \mathcal{N}, p \in \mathcal{P}$. Введем на множествах $\{\mathbf{x}_n : n \in \mathcal{N}\}$ и $\{\mathbf{x}_p : p \in \mathcal{P}\}$ отношения обобщенного доминирования $\succ_n$ и $\succ_p$. Объект $\mathbf{x}_n$ n-доминирует объект $\mathbf{x}_i$, если справедливы неравенства:

$$\mathbf{x}_n \succ_n \mathbf{x}_i, \quad \text{если} \quad x_{nj} \geq x_{ij} \quad \text{для всех} \quad j \in \mathcal{J}', \quad \text{или}$$

$$x_{nj} \geq x_{ij} \quad \text{для всех} \quad j \in \mathcal{J}' \setminus \{r, t\} \quad \text{и}$$

$$x_{nr} \geq x_{it}, \quad x_{nt} \geq x_{ir} \quad \text{при условии} \quad r \succ_j t \text{ и } x_{nr} > x_{nt}.$$

Аналогично, объект $\mathbf{x}_p$ p-доминирует $\mathbf{x}_k$, если справедливы неравенства:

$$\mathbf{x}_p \succ_p \mathbf{x}_k, \quad \text{если} \quad x_{pj} \leq x_{kj} \quad \text{для всех} \quad j \in \mathcal{J}', \quad \text{или}$$

$$x_{nj} \leq x_{kj} \quad \text{для всех} \quad j \in \mathcal{J}' \setminus \{r, t\} \quad \text{и}$$

$$x_{nr} \leq x_{kt}, \quad x_{nt} \leq x_{kr} \quad \text{при условии} \quad r \succ_j t \text{ и } x_{nr} < x_{nt}.$$

Будем считать, что объект не доминирует сам себя ни в одном из смыслов:

$$\mathbf{x} \not\succ_n \mathbf{x}, \quad \mathbf{x} \not\succ_p \mathbf{x}.$$

На рис. 5 приведен пример доминирования для случая двух признаков, первый из которых важнее второго. По осям отложены значения признаков, зелеными точками показаны объекты, которые доминируются объектами со значениями признаков (4;3) и (5;6) с учетом важности признаков. Желтым цветом закрашены расширенные области доминирования рассматриваемых объектов.

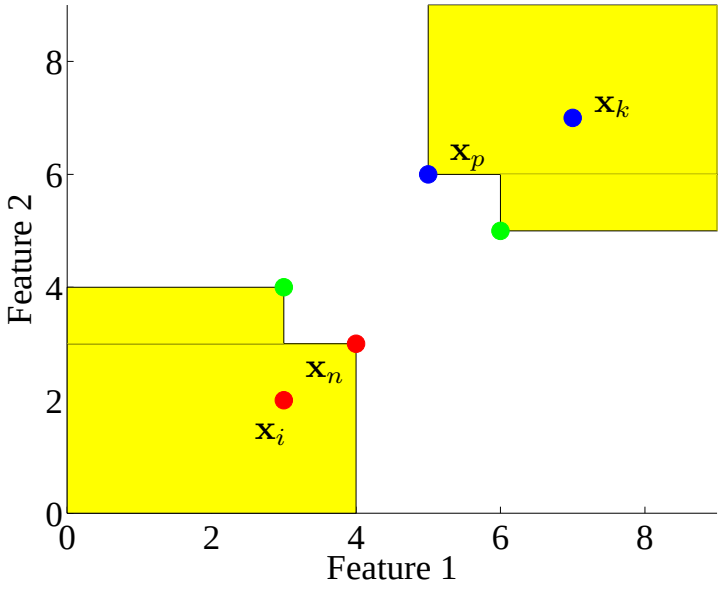
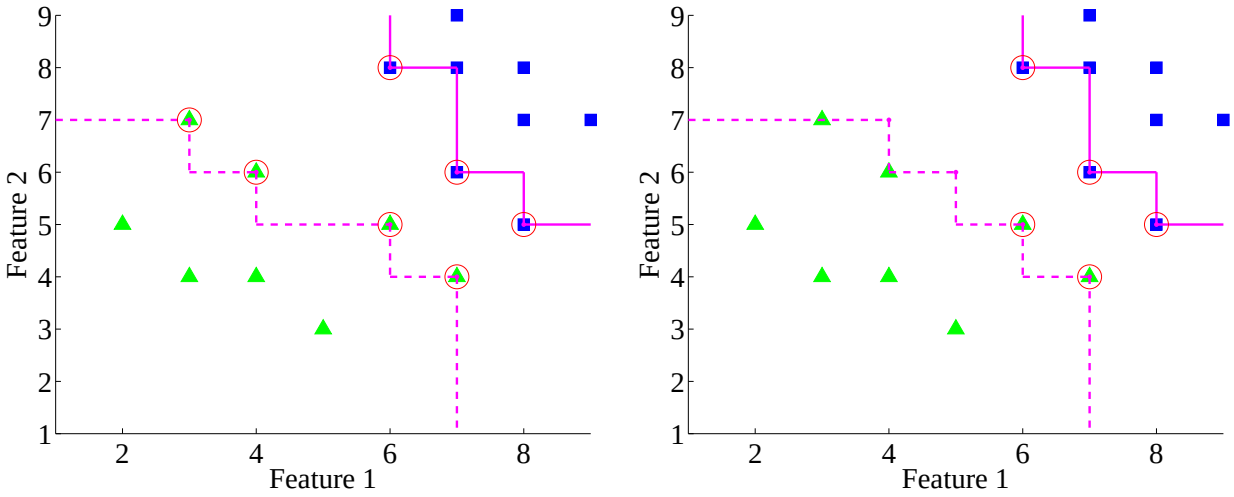


Рис. 5: Расширение областей доминирования при учете важности признаков

3.3 Построение Парето-оптимальных фронтов

Парето-оптимальный фронт POF_n — такое множество объектов $\mathbf{x}_n, n \in \mathcal{N}$, для каждого элемента которого $\mathbf{x}_n \in \mathrm{POF}_n$ не существует ни одного объекта $\mathbf{x}$, такого, что $\mathbf{x} \succ_n \mathbf{x}_n$. Парето-оптимальный фронт POF_p — такое множество объектов $\mathbf{x}_p, p \in \mathcal{P}$, для каждого элемента которого $\mathbf{x}_p \in \mathrm{POF}_p$ не существует объекта $\mathbf{x}$, такого, что $\mathbf{x} \succ_p \mathbf{x}_p$.



(a) Без учета важности признаков

(b) С учетом важности признаков

Рис. 6: Парето-оптимальные фронты

На рис. 6 показаны примеры Парето-оптимальных фронтов для случая двух признаков, значения которых отложены по осям графиков. Зелеными треугольниками и синими квадратами обозначены объекты разных классов. Граница класса, задаваемая n-фронтом, обозначена пунктирной линией, p-фронтом — сплошной. На рис. 6(a) изображены Парето-оптимальные фронты, соответствующие отношению доминирования без учета важности признаков. На рис. 6(b) изображены Парето-оптимальные фронты, соответствующие отношению доминирования с учетом важности признаков (первый признак важнее второго). Объекты выборки, вошедшие во фронты, обведены красными окружностями.

Как следует из этих рисунков, Парето-фронты, построенные с учетом важности признаков, содержат меньшее число объектов и доминируют более обширные области.

Далее во всех рассуждениях и выкладках будут использоваться отношения доминирования с учетом важности признаков.

3.4 Классификация для случая двух классов

Метка класса y ставится в соответсвие произвольному вектору $\mathbf{x}$ найденной функцией $f\colon \mathbf{x} \mapsto \hat{y}$. При этом, если найдутся некоторые векторы $\mathbf{x}_n$ или $\mathbf{x}_p$, где $n, p \in \hat{\mathcal{I}}$, которые находятся с $\mathbf{x}$ в отношении доминирования, то

$$f(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}_n \succ_n \mathbf{x}; \\ 1, & \mathbf{x}_p \succ_p \mathbf{x}. \end{cases} \tag{5}$$

Если таких элементов не найдется, то функция f доопределяется до множества $\mathbb{X}'$ (4) согласно правилу ближайшего множества POF:

$$f(\mathbf{x}) = f\left(\underset{\mathbf{x}' \in \overline{\mathrm{POF}}_n \cup \overline{\mathrm{POF}}_p}{\arg\min} \left(\rho\big(\mathbf{x}, \mathbf{x}'\big)\right) \right),$$

где множества $\overline{\mathrm{POF}}_n, \overline{\mathrm{POF}}_p$, включающие все точки границы области доминирования соответствующих Парето-оптимальных фронтов, включая объекты выборки, составляющие фронт, однозначно заданы ранее найденными множествами $\mathcal{N}, \mathcal{P}$ индексов элементов $\mathbf{x}_i \in \mathbb{X}'$. Функция ρ задана с помощью функции (3), примененной к меткам

значений признаков:

$$\rho(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} r(x_j, x'_j). \tag{6}$$

Таким образом, если не находится элементов, доминирующих объект $\mathbf{x}$, то $\mathbf{x}$ относится к тому классу, к Парето-оптимальному фронту которого он ближе.

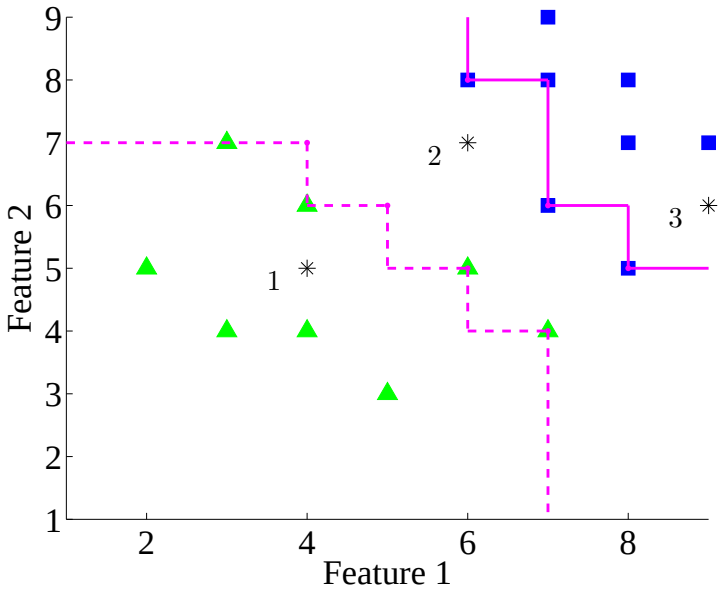


Рис. 7: Пример двухклассовой классификации

Таблица 2: Пример работы двухклассового классификатора

№	Объект $\mathbf{x}$	$f(\mathbf{x})$
1	(4,5)	0
2	(6,7)	1
3	(9,6)	1

На рис. 7 изображена синтетическая выборка, содержащая объекты двух классов. Объекты первого класса обозначены зелеными треугольниками, второго — синими квадратами. Объекты описаны двумя признаками, значения которых отложены по осям. Классифицируемые объекты на графике изображены черными звездочками. Результат работы классификатора f для этих объектов приведен в табл. 2, содержащей два столбца. В первом столбце координаты объектов, заданные значениями признаков, во втором — результаты работы классификатора. «0» во втором столбце

означает, что объект отнесен к первому классу, который обозначен зелеными треугольниками, «1» — ко второму классу, изображенному синими квадратами.

3.5 Приведение выборки к разделимой

Рассмотрим процедуру нахождения множества $\hat{\mathcal{I}}$, на котором функция $f\colon \mathbf{x} \mapsto \hat{y}$ монотонна. Разобьем множество индексов $\mathcal{I}$ объектов выборки $\mathfrak{D}$ на два подмножества $\mathcal{I} = \mathcal{N} \bigsqcup \mathcal{P}$ так, что $y_n = 0$, а $y_p = 1$, $n \in \mathcal{N}, p \in \mathcal{P}$. Рассмотрим мощность μ доминируемого объектом $\mathbf{x}_i$ множества объектов другого класса:

$$\mu(\mathbf{x}_i) = \begin{cases} \#\{\mathbf{x}_j \mid \mathbf{x}_i \succ_n \mathbf{x}_j, j \in \mathcal{P}\}, & \text{если } i \in \mathcal{N}; \\ \#\{\mathbf{x}_j \mid \mathbf{x}_i \succ_p \mathbf{x}_j, j \in \mathcal{N}\}, & \text{если } i \in \mathcal{P}, \end{cases}$$

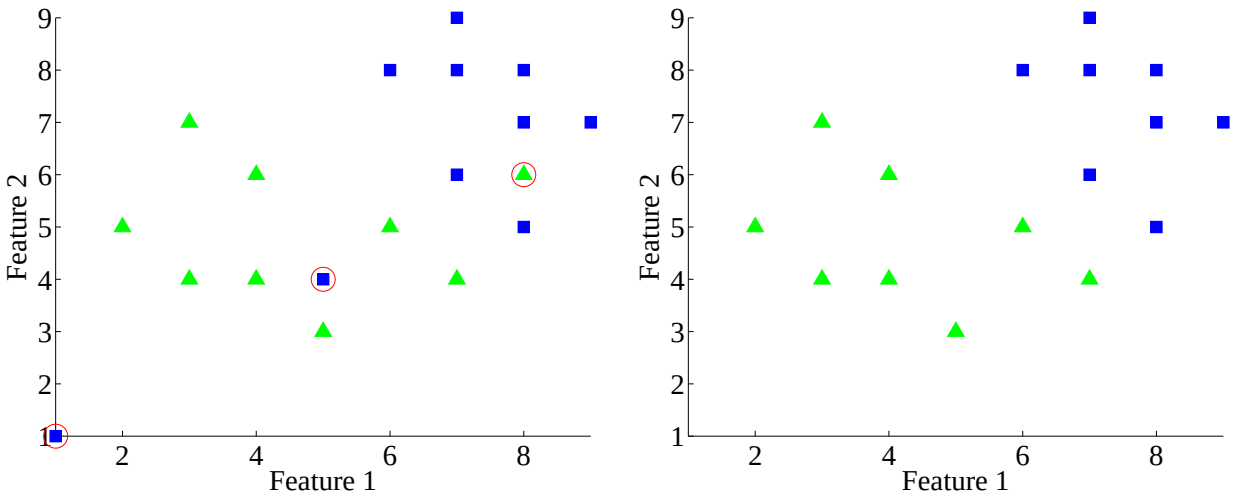
где знак # означает число элементов множества. Для нахождения множества $\hat{\mathcal{I}}$ проведем процедуру последовательного удаления объектов из выборки $\mathfrak{D}$.

Require: $\mathcal{I} = \mathcal{P} \bigcup \mathcal{N}$.

Ensure: $\hat{\mathcal{I}} = \hat{\mathcal{P}} \bigcup \hat{\mathcal{N}}$.

1: $\hat{\mathcal{I}} = \mathcal{I}$, $\hat{\mathcal{P}} = \mathcal{P}$, $\hat{\mathcal{N}} = \mathcal{N}$; {инициализация}
2: **while** в выборке есть объекты $\mathbf{x}_i$ с индексом $i \in \hat{\mathcal{I}}$ такие, что $\mu(\mathbf{x}_i) > 0$ **do**
3: $\quad \hat{i} = \arg\max_{i \in \hat{\mathcal{I}}} \mu(\mathbf{x}_i)$;
4: $\quad \hat{\mathcal{I}} = \hat{\mathcal{I}} \backslash \{\hat{i}\}$;
5: $\quad$ **if** $\hat{i} \in \hat{\mathcal{P}}$ **then**
6: $\quad\quad \hat{\mathcal{P}} = \hat{\mathcal{P}} \backslash \{\hat{i}\}$;
7: $\quad$ **if** $\hat{i} \in \hat{\mathcal{N}}$ **then**
8: $\quad\quad \hat{\mathcal{N}} = \hat{\mathcal{N}} \backslash \{\hat{i}\}$.

На рис. 8 изображена синтетическая выборка, объекты которой описываются двумя признаками. Выборка включает два класса, которые обозначены разными маркерами (зелеными треугольниками и синими квадратами). На рис. 8(a) изображена неразделимая выборка с дефектными объектами (1;1), (5;4) и (8;6), доминирующими объекты чужого класса. Эти объекты выделены красными окружностями. На рис. 8(b) показана разделимая выборка, полученная после применения описанной выше процедуры.



(a) С дефектными объектами

(b) Без дефектных объектов

Рис. 8: Синтетическая выборка

4 Монотонная классификация

4.1 Построение монотонного классификатора

Рассмотрим случай с более чем двумя классами; на множестве меток задано отношение линейного порядка. Пусть задано множество меток классов $\{1 \prec \cdots \prec u \prec v \prec \cdots \prec z\} = \mathbb{Z}$. Для каждой смежной пары классов u, v выше была определена монотонная функция $f_{uv}\colon \mathbf{x} \mapsto \hat{y} \in \{0, 1\}$, $\mathbf{x} \in \mathbb{X}$. Монотонный классификатор $\varphi(\mathbf{x}) = \varphi(f_{12}, \ldots, f_{(z-1)z})(\mathbf{x})$, функция $\varphi\colon \mathbb{X} \to \mathbb{Z}$, задан следующим образом:

$$\varphi(\mathbf{x}) = \begin{cases} \min\limits_{u \in \mathbb{Z}} \{u \mid f_{u,u+1}(\mathbf{x}) = 0\}, & \text{если } \{u \mid f_{u,u+1}(\mathbf{x}) = 0\} \neq \emptyset; \\ z, & \text{если } \{u \mid f_{u,u+1}(\mathbf{x}) = 0\} = \emptyset. \end{cases} \tag{7}$$

Функции f_{uv}, входящие в классификатор φ, строятся как описанные выше двухклассовые классификаторы. Метке «0» в множестве $\mathbb{Y} = \{0, 1\}$ соответсвуют классы с метками $y \leq u$, метке «1» — классы с метками $y \geq v$. Строится множество индексов объектов $\hat{\mathcal{I}}$, соответсвующее разделимой выборке $\hat{\mathfrak{D}}$. Парето-оптимальные фронты строятся для множеств $\{\mathbf{x}_n\colon n \in \mathcal{N}_u\}$ и $\{\mathbf{x}_p\colon p \in \mathcal{P}_v\}$, где множества индексов $\mathcal{N}_u, \mathcal{P}_v$ определяются следующим образом

$$\begin{aligned} n \in \mathcal{N}, &\quad \text{если} \quad y_n \leqslant u \quad \text{и} \\ p \in \mathcal{P}, &\quad \text{если} \quad v \leqslant y_p. \end{aligned} \tag{8}$$

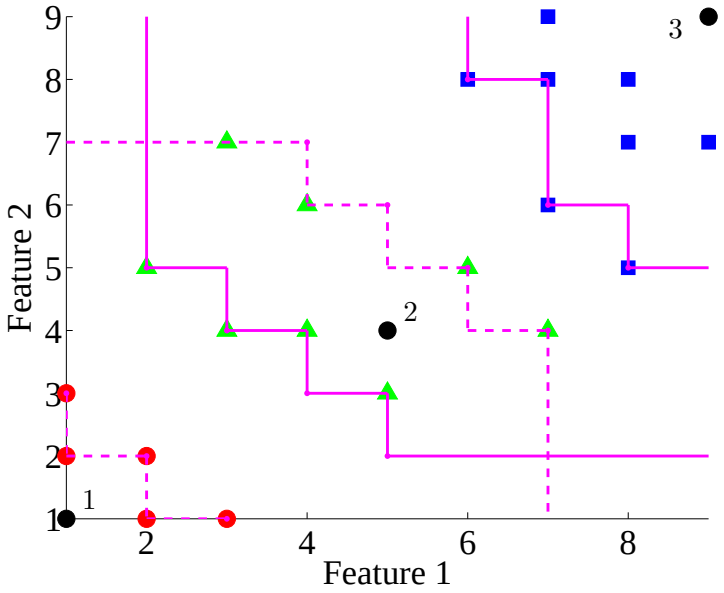


Рис. 9: Парето-фронты, первый признак важнее второго

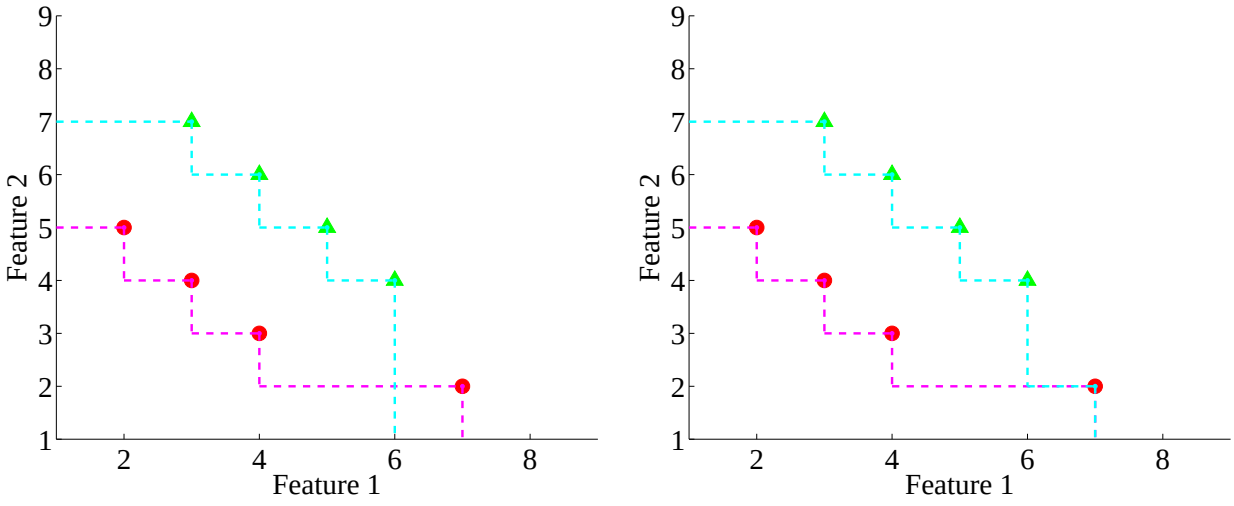
На рис. 9 изображена синтетическая выборка, содержащая объекты трех классов. Значения двух признаков, которыми описываются объекты, отложены по осям графика, объекты, принадлежащие разным классам, отмечены красными кружками, зелеными треугольниками и синими квадратами. Показаны построенные с учетом важности признаков фронты. n-фронты обозначены пунктирной линией, p-фронты — сплошной. Объекты выборки, вошедшие во фронты, отмечены светлыми точками. Классифицируемые объекты отмечены черными кружками. Пример результата работы набора функций f_{12}, f_{23}, входящих в классификатор φ, для выборки на рис. 9, выглядит как табл. 3. Первый столбец таблицы содержит координаты объектов, заданные значениями признаков, второй и третий столбцы — результаты работы двухклассовых классификаторов для смежных первого и второго, второго и третьего классов соответственно на представленных объектах. «0» во втором столбце означает, что объект был отнесен к первому классу классификатором f_{12}, «1» — ко второму классу этим же классификатором. «0» в третьем столбце означает, что объект был отнесен ко второму классу классификатором f_{23}, «1» — к третьему классу классификатором f_{23}. Последний столбец содержит результаты монотонной классификации объектов. Значения в этом столбце соответствуют номеру класса, к которому в итоге был отнесен объект.

Функция ошибки при монотонной классификации задана (2).

Таблица 3: Пример работы монотонного классификатора

№	Объект $\mathbf{x}$	$f_{12}(\mathbf{x})$	$f_{23}(\mathbf{x})$	$\varphi(\mathbf{x})$
1	(1,1)	0	0	1
2	(5,4)	1	0	2
3	(9,9)	1	1	3

4.2 Доопределение Парето-оптимальных фронтов при монотонной классификации



(a) Без доопределения

(b) С доопределением

Рис. 10: Общий объект для двух n-фронтов

Как следует из формулы (8), при построении фронтов между классами с метками u и $u+1$ используются объекты классов с метками $1,\ldots,u$ для построения n-фронта для класса u и объекты классов с метками $u+1,\ldots,z$ для построения p-фронта для класса $u+1$. Поэтому одни и те же объекты могут попадать во фронты для разных классов, доопределяя их, и фронт одного класса может содержать объекты нескольких классов. На рис. 10 приведен фрагмент синтетической выборки, содержащей объекты трех классов. На графике изображены только объекты первого (красные кружки) и второго (зеленые треугольники) классов, иллюстрирующие ситуацию, когда объект с координатами (7;2) из первого класса попадает в n-фронты

первого и второго классов.

Таким образом, n-фронт доопределяется объектами, принадлежащими классам с метками, не превосходящими метку класса, для которого этот фронт строится; p-фронт доопределяется аналогичным образом объектами классов с метками, превосходящими метку класса,для которого строится фронт.

4.3 Допустимые классификаторы

Классификатор φ (7) будем называть *допустимым*, если для всех входящих в него функций f_{uv} соблюдается условие транзитивности:

$$\begin{cases} \text{если} \quad f_{uv}(\mathbf{x}) = 0, \quad \text{то} \quad f_{(u+s)(v+s)} = 0 \quad \text{для всех} \quad s\colon (v+s) \leqslant z, \\ \text{если} \quad f_{uv}(\mathbf{x}) = 1, \quad \text{то} \quad f_{(u-s)(v-s)} = 1 \quad \text{для всех} \quad s\colon (u-s) \geqslant 1. \end{cases} \tag{9}$$

Обозначим $\overline{\mathrm{POF}_n(u)}$ и $\overline{\mathrm{POF}_p(u+1)}$, $u = 1, \ldots, z-1$, все точки границы области доминирования соответствующих Парето-оптимальных фронтов, включая объекты исходной выборки, составляющие фронт.

Определение 1 *Парето-оптимальные фронты* $POF_n(u)$ *и* $POF_p(u+1)$, $u = 1, \ldots, z-1$ *называются непересекающимися* $POF_n(u) \cap POF_p(u+1) = \emptyset$, *если границы их областей доминирования* $\overline{POF_n(u)}$ *и* $\overline{POF_p(u+1)}$ *не имеют общих точек.*

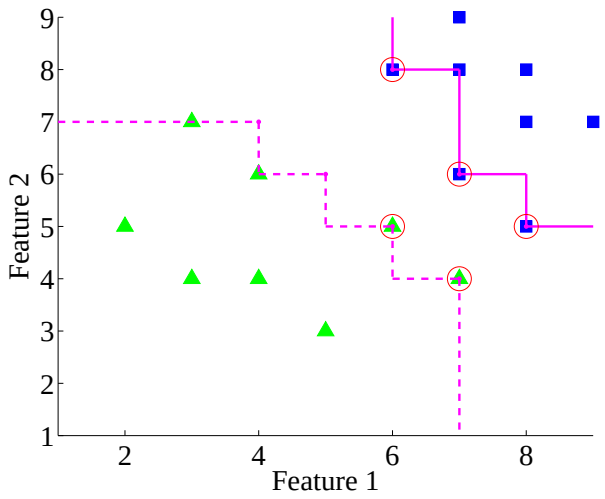


Рис. 11: Пример непересекающихся Парето-оптимальных фронтов

На рис. 11 изображена синтетическая выборка объектов двух классов (зеленые треугольники и синие квадраты), описанных двумя признаками. Построены непересекающиеся фронты, с учетом важности признаков (первый признак важнее второго).

Теорема 1 *Непересечения Парето-оптимальных фронтов* $POF_n(u) \cap POF_p(u+1) = \emptyset$, $u = 1, \ldots, z-1$ *достаточно для выполнения отношения транзитивности (9) для любого классифицируемого объекта.*

Доказательство. Проведем доказательство для $z = 3$. Для большего числа классов обобщение строится аналогично приведенному ниже доказательству.

Предположим, что выполнено условие непересечения Парето-оптимальных фронтов:

$$\mathrm{POF}_n(u) \quad \cap \quad \mathrm{POF}_p(u+1) = \emptyset, \quad u = 1, 2;$$

и существует объект $\mathbf{x}$, на котором нарушается отношение транзитивности:

$$f_{12}(\mathbf{x}) = 0, \quad f_{23}(\mathbf{x}) = 1$$

(случай $f_{12}(\mathbf{x}) = 1, \quad f_{23}(\mathbf{x}) = 0$ рассматривается путем зеркального отражения).

Результат $f_{12}(\mathbf{x}) = 0$ может быть получен в двух случаях.

1) $\exists \mathbf{y} \in \mathrm{POF}_n(1) : \quad \mathbf{y} \succ_n \mathbf{x}$. Если $\mathbf{y} \in \mathrm{POF}_n(2)$, то из этого автоматически вытекает, что $f_{23}(\mathbf{x}) = 0$, что противоречит предположению $f_{23}(\mathbf{x}) = 1$. Если $\mathbf{y} \notin \mathrm{POF}_n(2)$, то $\exists \mathbf{w} \in \mathrm{POF}_n(2) : \quad \mathbf{w} \succ_n \mathbf{y}$. Тогда получаем цепочку отношений доминирования $\mathbf{w} \succ_n \mathbf{y} \succ_n \mathbf{x}$, из которой заключаем, что $\mathbf{w} \succ_n \mathbf{x}$. И опять приходим к тому, что $f_{23}(\mathbf{x}) = 0$.

2) $\mathbf{x}$ не доминируется фронтами $\mathrm{POF}_n(1)$ и $\mathrm{POF}_p(2)$. В этом случае $\exists \mathbf{y}_0 \in \overline{\mathrm{POF}_n(1)}$:

$$\mathbf{y}_0 = \underset{\mathbf{y} \in \overline{\mathrm{POF}_n(1)} \cup \overline{\mathrm{POF}_p(2)}}{\arg\min} \rho(\mathbf{x}, \mathbf{y}),$$

где ρ — расстояние (6).

Мы предположили, что $f_{23}(\mathbf{x}) = 1$. Это возможно в двух случаях.

а) $\exists \mathbf{t} \in \mathrm{POF}_p(3) : \quad \mathbf{t} \succ_p \mathbf{x}$. Этот объект $\mathbf{t}$ не может лежать в $\mathrm{POF}_p(2)$, так как $\mathbf{x}$ не доминируется фронтом $\mathrm{POF}_p(2)$. Но в таком случае $\exists \mathbf{t}_0 \in \mathrm{POF}_p(2) : \quad \mathbf{t}_0 \succ_p \mathbf{t}$.

Тогда получаем цепочку $\mathbf{t}_0 \succ_p \mathbf{t} \succ_p \mathbf{x}$, из которой вытекает $\mathbf{t}_0 \succ_p \mathbf{x}$, что противоречит предположению о том, что $\mathbf{x}$ не доминируется фронтом $\mathrm{POF}_p(2)$.

б) $\mathbf{x}$ не доминируется фронтами $\mathrm{POF}_n(2)$ и $\mathrm{POF}_p(3)$. В таком случае $\mathbf{x}$ не доминируется ни одним из фронтов $\mathrm{POF}_n(u), \quad \mathrm{POF}_p(u+1), \quad u = 1, 2$.

Заметим, что существует объект $\mathbf{y}_1 \in \mathrm{POF}_n(1)$, граница области доминирования которого содержит точку $\mathbf{y}_0$, ближайшую к $\mathbf{x}$ в смысле расстояния (6). При построении $\mathrm{POF}_n(2)$ объект $\mathbf{y}_1$ может войти в этот фронт. Тогда расстояние от $\mathbf{x}$ до ближайшей к нему точки фронта $\mathrm{POF}_n(2)$ не больше, чем расстояние от $\mathbf{x}$ до ближайшей к нему точки $\mathbf{y}_0$ фронта $\mathrm{POF}_n(1)$. Объект $\mathbf{y}_1$ может не войти в $\mathrm{POF}_n(2)$. Тогда существует объект $\mathbf{y}_2 \in \mathrm{POF}_n(2) : \quad \mathbf{y}_2 \succ_n \mathbf{y}_1$, но при этом $\mathbf{y}_2 \not\succ_n \mathbf{x}$, т. к. $\mathbf{x}$ не доминируется ни одним из фронтов. И в таком случае расстояние от $\mathbf{x}$ до ближайшей к нему точки фронта $\mathrm{POF}_n(2)$ не больше, чем расстояние от $\mathbf{x}$ до ближайшей к нему точки $\mathbf{y}_0$ фронта $\mathrm{POF}_n(1)$.

Аналогичное рассуждение проводится и для пары фронтов $\mathrm{POF}_p(2), \quad \mathrm{POF}_p(3)$. Существует объект $\mathbf{w}_1 \in \mathrm{POF}_p(2)$, граница области доминирования которого содержит точку $\mathbf{w}_0$, ближайшую к $\mathbf{x}$ в смысле расстояния (6). При построении $\mathrm{POF}_p(3)$ в независимости от того, вошел ли объект $\mathbf{w}_1$ в этот фронт, расстояние от $\mathbf{x}$ до ближайшей к нему точки фронта $\mathrm{POF}_p(3)$ не меньше, чем расстояние от $\mathbf{x}$ до ближайшей к нему точки $\mathbf{w}_0$ фронта $\mathrm{POF}_p(2)$.

В итоге имеем, что $\mathrm{POF}_n(2)$ не дальше от $\mathbf{x}$, чем $\mathrm{POF}_n(1)$, а $\mathrm{POF}_p(3)$ не ближе к $\mathbf{x}$, чем $\mathrm{POF}_p(2)$. Так как из $f_{12}(\mathbf{x}) = 0$ следует, что $\mathbf{x}$ ближе к $\mathrm{POF}_n(1)$, чем к $\mathrm{POF}_p(2)$, то $\mathbf{x}$ ближе к $\mathrm{POF}_n(2)$, чем к $\mathrm{POF}_p(3)$, что противоречит условию $f_{23}(\mathbf{x}) = 1$.

■

Поскольку в работе для построения Парето-оптимальных фронтов используются разделимые выборки, все построенные фронты являются непересекающимися. Поэтому построенный монотонный классификатор (7) допустим и для любого классифицируемого объекта выполняется условие транзитивности (9).

5 Вычислительный эксперимент

5.1 Заполнение пропусков в данных

Проведен эксперимент по выбору способа заполнения пропусков. Для этого использовались значения признаков у объектов выборки, ближайших к объекту с пропуском. Близость объектов $\mathbf{x}_i$ и $\mathbf{x}_j$ определялась расстоянием (6). Использовалась схема Leave-One-Out. На каждой итерации из обучения исключался один из объектов, который использовался для контроля, затем вычислялось значение функции ошибки.

$$\text{LOO} = \frac{1}{m}\sum_{i=1}^{m} r(y_i, \varphi_i(\mathbf{x}_i)), \tag{10}$$

где φ_i — монотонный классификатор (7), построенный с помощью выборки $\mathfrak{D}_i = \{(\mathbf{x}_j, y_j)\}$, $j \in \mathcal{I} = \{1, \ldots, i-1, i+1, \ldots, m\}$; r задана в (3).

Обозначим α способ заполнения пропусков. Оптимальный способ заполнения пропусков $\hat{\alpha}$ определяется как

$$\hat{\alpha} = \arg\min_{\alpha} \text{LOO}^{\alpha} = \frac{1}{m}\sum_{i=1}^{m} r(y_i, \varphi_i^{\alpha}(\mathbf{x}_i)).$$

Рассматривались способы заполнения пропусков средним значением признака у трех объектов, ближайших к объекту с пропуском и средним значением признака у трех ближайших объектов из класса, которому принадлежит объект с пропуском. Также были рассмотрены варианты без пересчета начальной матрицы попарных расстояний между объектами и с ее пересчетом после каждого заполненного пропуска. Пропуски в контрольных объектах заполнялись по трем ближайшим объектам из обучающей выборки.

Вычисление расстояния между объектами с пропусками. Использовалась модификация формулы (6). Для пары объектов $\mathbf{x}_i$, $\mathbf{x}_k$ по каждому признаку $j = 1, \ldots, d$ рассматривались следующие варианты, где функция r определена в (3).

Нет пропусков. В сумму входит слагаемое $r(x_{ij}, x_{kj})$.

Пропуск в объекте $\mathbf{x}_i$. Рассматриваются все допустимые значения $1, \ldots, l_j$ пропущенного элемента, для каждого варианта вычисляется $r(1, x_{kj}), \ldots, r(l_j, x_{kj})$. В

сумму включается мода полученного распределения, если мода не существует, то медиана. Для пропущенного значения объекта $\mathbf{x}_k$ аналогично.

Пропуск в обоих объектах. Вычисляются абсолютные значения разностей для всех допустимых пар с учетом порядка. В сумму включается мода полученного распределения, если мода не существует, то медиана.

Результаты представлены в табл. 4. В первом столбце таблицы — названия способов заполнения пропусков, во втором — значения LOO, полученные для этих способов заполнения. По наименьшему значению функции потерь выбран способ заполнения пропусков средним значением признака у трех ближайших объектов из класса, которому принадлежит объект с пропуском, с пересчетом начальной матрицы попарных расстояний между объектами.

Таблица 4: Способы заполнения пропусков

Способ заполнения	LOO
Три ближайших соседа без пересчета расстояний	0.6961
Три ближайших соседа с пересчетом расстояний	0.6275
Три ближайших соседа из своего класса без пересчета расстояний	0.5588
Три ближайших соседа из своего класса с пересчетом расстояний	0.6569

5.2 Сравнение алгоритмов

В ходе вычислительного эксперимента сравнивались монотонные классификаторы $\varphi_{\mathcal{A}_1}, \ldots, \varphi_{\mathcal{A}_5}$, построенные с помощью различных поднаборов признаков, монотонный классификатор φ, построенный по результатам работы первых пяти классификаторов, алгоритм классификации с помощью решающих деревьев, алгоритм криволинейной регрессии, алгоритмы с использованием конусов и копул. Все модели сравнивались по трем показателям: средняя ошибка на обучающей выборке, LOO и экспериментальное время построения модели.

Для вычисления средней ошибки на обучающей выборке алгоритму с настроенными параметрами на вход подавалась выборка, с помощью которой проходило обучение. Ошибка вычислялась по формуле (2). Алгоритм вычисления LOO описан

в пункте «Заполнение пропусков в данных» (10). Экспериментальное время оценивалось как время, необходимое для настройки параметров модели.

Результаты эксперимента представлены в табл. 5.

Таблица 5: Сравнение алгоритмов классификации

Алгоритм	Средняя ошибка на обучении	LOO	Время построения модели, сек
POF	0.2157	0.5588	2.1251
Решающие деревья	0.2451	0.6863	0.4154
Криволинейная регрессия [33]	0.57	0.71	3.6
Конусы [35]	0.29	0.58	1.2
Копулы [34]	0.57	0.61	0.25

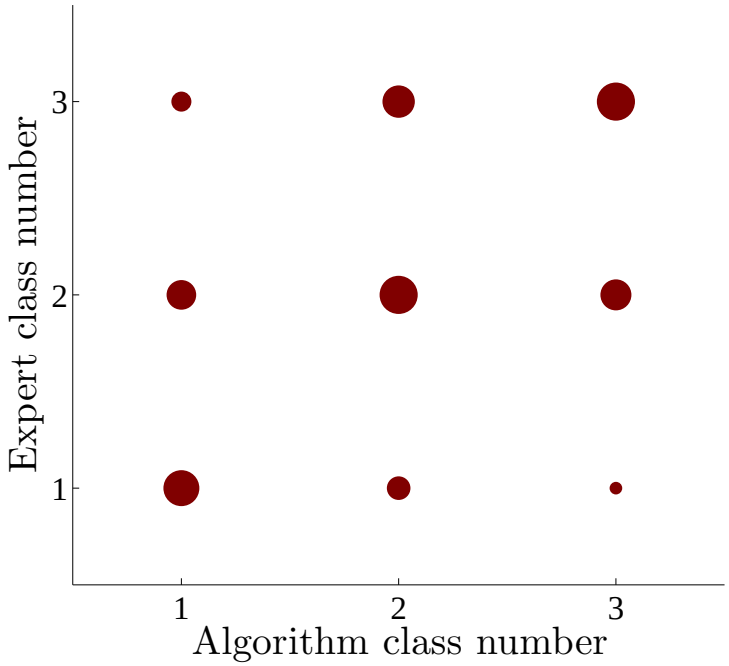


Рис. 12: Сравнение вычисленных и экспертных категорий риска

На рис. 12 сравниваются категории риска, вычисленные при тестировании алгоритма с помощью LOO, и экспертные категории риска. По оси абсцисс отложены метки вычисленных категорий, по оси ординат — экспертных. Размер точки соответствует количеству видов, попавших на пересечение данной вычисленной категории

риска и экспертной категории. Как видно на рисунке, для значительной части объектов вычисленная и экспертная категории совпадают, для меньшей части номера категорий отличаются на один, еще для более незначительной части — на два. В табл. 6 представлены те же результаты с указанием видов для каждой пары вычисленной и экспертной категорий. Столбцы таблицы соответствуют вычисленным категориям риска, строки — экспертным.

5.3 Работа алгоритма на различных наборах данных

Для тестирования алгоритма помимо предоставленных данных о редких видах были использованы наборы Cars и CPU из репозитория UCI. Для набора CPU задача восстановления регрессии была преобразована в задачу классификации путем разбиения объектов выборки на четыре равномощных класса по значениям целевой переменной. Для контроля качества использовались функционалы LOO (10) и 10-fold. Для вычисления последнего выборка случайным образом разбивается на 10 одинаковых по мощности частей. Алгоритм обучается на девяти десятых выборки, одна десятая используется для контроля. Процедура повторяется для каждой доли выборки, результаты усредняются. Результаты эксперимента представлены в табл. 7, в последнем столбце которой представлены результаты алгоритма, предложенного в работе [36]. Прочерк в ячейке таблицы означает, что данный эксперимент не проводился. Для наборов Cars и CPU иерархия и важность признаков не вводилась ввиду отсутствия экспертной информации и малого количества признаков.

5.4 Выделение опорных объектов

Выделение опорных объектов классов производилось для предоставления экспертам информации о наиболее типичных видах для каждой категории риска. Опорные объекты находились при помощи алгоритма STOLP [37]. Работа алгоритма зависит от трех параметров: ширина окна, допустимое количество ошибок при классификации и порог отсеивания шумовых объектов.

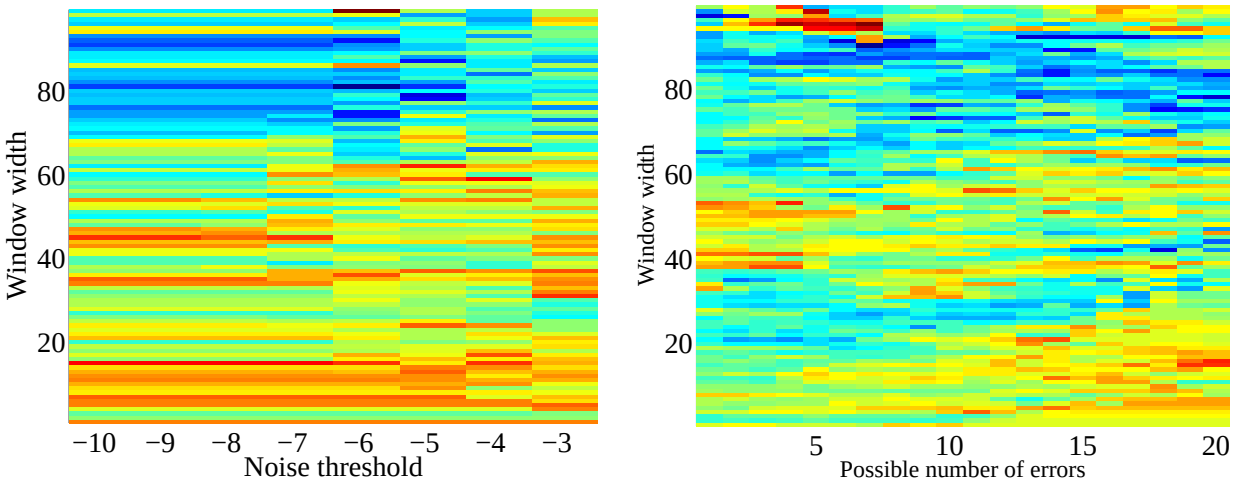
На рис. 13 изображены значения функционала LOO (10) для различных значений параметров. Значения LOO обозначены цветом ячеек. Большим значениям соответствуют оттенки красного и желтого, меньшим — оттенки синего. На основе

Таблица 6: Сравнение алгоритмов классификации

	Вычисленные категории риска		
Class labels	1	2	3
1	Азовская белуга Схизофрагма гортензиевидная Миякея цельнолистная Морская минога Калуга Азовская белуга Сахалинский осетр Амурский осетр Шип Кумжа черноморский подвид проходная форма Обыкновенный таймень Желтощек Черный амур Мелкочешуйный желтопер Кильдинская треска Стерх белый журавль обская западная популяция	Сахалинский осетр Береза Максимовича Гнездовка уссурийская Горал Дальневосточный леопард Стерлядь Кумжа каспийский подвид проходная форма Ленок Нельма подвид белорыбица Черный амурский леш Савка	Кильдинская треска Нельма подвид нельма Днепровский усач Средиземноморская черепаха Японский журавль
2	Сибирский осетр западносибирский обский подвид Сахалинский таймень Сахалинский таймень Русская быстрянка Китайский окунь или ауха Гадюка Казнакова Гюрза	Сибирский осетр байкальский подвид Амурский тигр Уссурийский пятнистый олень Украинская минога Сибирский осетр байкальский подвид Волжская сельдь Озерный лосось пресноводная форма атлантического лосося семги Кумжа беломорскобалтийский подвид бассейн Балтийского моря Кумжа подвид эйзенамская форель Арктический голец популяция Забайкалья Сиг волховский подвид Переславская ряпушка Европейский хариус Вырезуб подвид кутум Азовочерноморская шемая Сом Солдатова Обыкновенный подкаменщик Дальневосточная черепаха	Волховский сиг Водяной орех чилим Луговик Турчанинова Остролодочник тодомоширский Цетрария степная Каспийская минога Камышовая жаба Уссурийский когтистый тритон Эскулапов полоз Гадюка Динника Русская выхухоль Пискулька Лобария легочная
3	Куликлопатень Солнцецвет арктический	Белый медведь лаптевская популяция Кизильник киноварно-красный Маннагеттея Гуммеля Осмунда японская Чистоус японский Микижа проходная форма камчатская семга Сиг баунтовский подвид Предкавказская щиповка Закавказский полоз Краснопоясный одинодон Средняя ящерица Краснозобая казарка Стерх белый журавль якутская восточная популяция	Длинноперая палия Световидова Желтозобик Мелколепестник сложноцветный Сердечник пурпурный Калопанакс семилопастный Омфалина гудзонская Малоротая палия Длинноперая палия Световидова Карликовый валек Берш популяция бассейна реки Урал Дальнвосточный сцинк Западный удавчик Кошачья змея Пискливый гекконочик Полосатый полоз Японский полоз Ящурка Пржевальского Кавказский тетерев

Таблица 7: Работа алгоритма на различных наборах данных

Данные	Количество объектов	Количество признаков	Количество классов	POF LOO	POF 10-fold	LPRules 10-fold
Красная книга	102	102	3	0.5588	—	—
Cars	1728	6	4	0.3553	0.1933	0.03
CPU	209	6	4	0.6411	0.4833	0.073



(a) Фиксировано число ошибок

(b) Фиксирован порог

Рис. 13: LOO для различных значений параметров

этих рисунков производился выбор стартовой точки для алгоритма численной оптимизации. Ширина окна была выбрана 90, допустимое количество ошибок 8, порог отсеивания шума -7. В результате минимизации LOO получены значения параметров ширина окна 90, допустимое количество ошибок 10, порог отсеивания шума -7. Значение функционала LOO = 0.6667.

Заключение

Построен алгоритм многоклассовой монотонной классификации объектов, описанных в ранговых шкалах. Алгоритм использует Парето-оптимальные фронты двух типов, построенные на основе отношения доминирования с учетом экспертной информации о важности признаков и включает два уровня классификации, учитывающие иерархию признаков. Предложен способ заполнения пропущенных значений в признаковых описаниях объектов. Проведено сравнение предлагаемого алгоритма с рядом других алгоритмов, которое показало, что предлагаемый алгоритм при использовании всех признаков и информации об их структуре дает более надежные результаты классификации.

Публикации по теме

1. Медведникова М. М. Использование метода главных компонент при построении интегральных индикаторов // Машинное обучение и анализ данных. 2012. № 3. С. 292–304.

2. Кузнецов М. П., Стрижов В. В., Медведникова М. М. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах // НТВСПбГПУ. 2012. № 5. С. 92–94.

3. Медведникова М. М. Стрижов, В. В., Кузнецов М. П. Алгоритм многоклассовой монотонной Парето-классификации с выбором признаков // Известия ТулГУ. Естественные науки. 2012. № 3. С. 132–141.

4. Медведникова М. М., Стрижов В. В. Построение интегрального индикатора качества научных публикаций методами ко-кластеризации // Известия ТулГУ.

Естественные науки. 2013. № 1.

5. Вальков А. С., Кожанов Е. М., Медведникова М. М., Хусаинов Ф. И. Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных. 2012. № 4. С. 448–465.

Список литературы

[1] Красная книга Российской Федерации. М.: Институт проблем экологии и эволюции имени А. Н. Северцова РАН // Под ред. В. И. Данилов-Данильян и др. http://www.sevin.ru/redbook/ (31.07.2012).

[2] Красная книга Российской Федерации (животные) // М: АСТ Астрель, 2001.

[3] Лбов Г. С. Выбор эффективной системы зависимых признаков // Вычислительные системы. 1965. Т. 19. С. 21–34.

[4] Рао С. Линейные и статистические методы и их примениния // Наука. 1968. С. 530–533.

[5] Park J. S., Chen M.-S., Yu P. S. An effective hash-based algorithm for mining association rules // SIGMOD Rec., ACM. 1995. Vol. 24. P. 175–186.

[6] Ногин В. Границы применимости распространенных методов скаляризации при решении задач многокритериального выбора // Межвуз. сб. научн. тр.–Саранск: Изд-во Мордовского ун-та. 2004. С. 59–68.

[7] Liu T.-Y., Joachims T., Li H., Zhai C. Introduction to special issue on learning to rank for information retrieval // Information Retrieval, Springer Netherlands. 2010. Vol. 13. P. 197–200.

[8] Cheng W., Rademaker M., De Baets B., Hullermeier E. Predicting partial orders: ranking with abstention // Machine Learning and Knowledge Discovery in Databases, Springer. 2010. P. 215–230.

[9] Cossock D., Zhang T., Lugosi G., Simon H. Subset Ranking Using Regression // Learning Theory, Springer Berlin Heidelberg. 2006. Vol. 4005. P. 605–619.

[10] Yue Y., Finley T., Radlinski F., Joachims T. A support vector method for optimizing average precision // Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM. 2007. P. 271–278.

[11] Freund Y., Iyer R., Schapire, R. E., Singer Y. An efficient boosting algorithm for combining preferences // J. Mach. Learn. Res., JMLR.org. 2003. Vol. 4. P. 933–969.

[12] Schmidt G. Relational mathematics // Cambridge University Press. 2010. 132 p.

[13] Agresti A. An introduction to categorical data analysis // Wiley-Interscience. 2007. 423 p.

[14] Орлов А. Нечисловая статистика // М.: МЗ-Пресс. 2004. 436 с.

[15] Doyle J. Prospects for preferences // Computational Intelligence, Wiley Online Library. 2004. Vol. 20 P. 111–136.

[16] Furnkranz J., Hullermeier E. Pairwise preference learning and ranking // Machine Learning: ECML 2003, Springer. 2003. P. 145–156.

[17] Har-Peled S., Roth D., Zimak D. Constraint Classification for Multiclass Classification and Ranking // NIPS. 2003. P. 785–792.

[18] Hullermeier E., Furnkranz J. Learning from label preferences // Discovery Science. 2011. P. 2–17.

[19] Hullermeier E., Furnkranz J. Learning preference models from data: On the problem of label ranking and its variants // Preferences and Similarities, Springer. 2008. P. 283–304.

[20] Hullermeier E., Furnkranz J. Comparison of ranking procedures in pairwise preference learning // Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-04). Perugia, Italy. 2004.

[21] Hullermeier E., Furnkranz J., Cheng W., Brinker K. Label ranking by learning pairwise preferences // Artificial Intelligence, Elsevier. 2008. Vol. 172. P. 1897–1916.

[22] Xia F., Liu T.-Y., Wang J., Zhang W., Li H. Listwise approach to learning to rank: theory and algorithm // Proceedings of the 25th international conference on Machine learning, ACM. 2008. P. 1192–1199.

[23] Ailon N. An Active Learning Algorithm for Ranking from Pairwise Preferences with an Almost Optimal Query Complexity // Machine Learning Research. 2012. Vol. 13(Jan). P. 137–164.

[24] Baehrens D., Schroeter T. How to Explain Individual Classification Decisions // Machine Learning Research. 2010. Vol. 11(Jun). P. 1803–1831.

[25] Cour T., Ben Sapp B. T. Learning from Partial Labels // Machine Learning Research. 2011. Vol. 12(May). P. 1501–1536.

[26] Подиновский В. В., Ногин В. Д. Парето-оптимальные решения многокритериальных задач // М.: Наука. 1982.

[27] Ногин В. Д. Логическое обоснование принципа Эджворта-Парето // Журнал вычислительной математики и математической физики, Российская академия наук, Отделение математических наук. 2002. Т. 42. С. 951–957.

[28] Ногин В. Д. Сужение множества Парето на основе информации о предпочтениях ЛПР точечно-множественного типа // Искусственный интеллект и принятие решений, Институт системного анализа РАН. 2009. С. 5–16.

[29] Ногин В. Д. Проблема сужения множества Парето: подходы к решению // Искусственный интеллект и принятие решений, Институт системного анализа РАН. 2008. С. 98–112.

[30] Подиновский В. В. Введение в теорию важности критериев // М.: Физматлит. 2007. 64 с.

[31] Медведникова М. М., Стрижов В. В., Кузнецов М. П. Алгоритм многоклассовой монотонной Парето-классификации с выбором признаков // Известия Тульского государственного университета. Естественные науки. 2012. № 3. С. 132–141.

[32] Законодательство в сфере охраны животного и растительного мира: Российская Федерация http://oopt.aari.ru/rbdata/900 (31.07.2012).

[33] Кузнецов М. П., Стрижов В. В., Медведникова М. М. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах // Научно-технический вестник СПбГПУ. Информатика. Телекоммуникации. Управление. 2012. № 5. С. 92–94.

[34] Кузнецов М. П. Построение интегрального индикатора в ранговых шкалах с использованием копул для анализа совместного распределения критериев // Машинное обучение и анализ данных. 2012. Т. 1. №4. С. 411–419.

[35] Кузнецов М. П., Стрижов В. В. Построение интегрального индикатора с использованием ранговой матрицы описаний // Доклады 9-й международной конференции по интеллектуализации обработки информации ИОИ-9. 2012. С. 130–132.

[36] Kotlowski W., Slowinski, R. Rule learning with monotonicity constraints // Proceedings of the 26th Annual International Conference on Machine Learning. 2009. P. 537–544.

[37] Загоруйко Н. Г. Прикладные методы анализа данных и знаний // Новосибирск: ИМ СО РАН. 1999.