Построение интегральных индикаторов по частично упорядоченным множествам экспертных оценок

Медведникова Мария

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва,
2013 г.

Цель работы

Задача

Требуется создать интегральный индикатор для определение категории риска редкого вида из Красной книги РФ. Задача поставлена экспертами Министерства природных ресурсов и экологии и WWF.

Используемые категории

1. находящиеся под угрозой исчезновения;
2. сокращающиеся в численности;
3. редкие.

Требования к модели

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

Существующие решения

Интегральные индикаторы

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. Список МСОП http://www.redbook.ru/msop.htm

Базовые публикации

1. Подиновский В. В. Введение в теорию важности критериев // М.: Физматлит. 2007. 64 с.
2. 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.
3. Стрижов В.В. Уточнение экспертных оценок, выставленных в ранговых шкалах, с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов, 2011, Т. 77(7). С. 72–78.

Входные данные

Фрагмент анкеты для описания вида экспертом

Вид: **русская выхухоль**

Критерий	Состояние	Тенденция изменения
Численность	3 – высокая; **2** – низкая; 1 – критически низкая	4 – растет; 3 – стабильна; **2** – медленно снижается; 1 – быстро снижается
Популяционная структура вида	2 – сложная; **1** – простая	**2** – стабильна; 1 – исчезают локальные популяции

На множестве признаков введен частичный порядок

Постановка задачи

Дано

множество пар $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$.

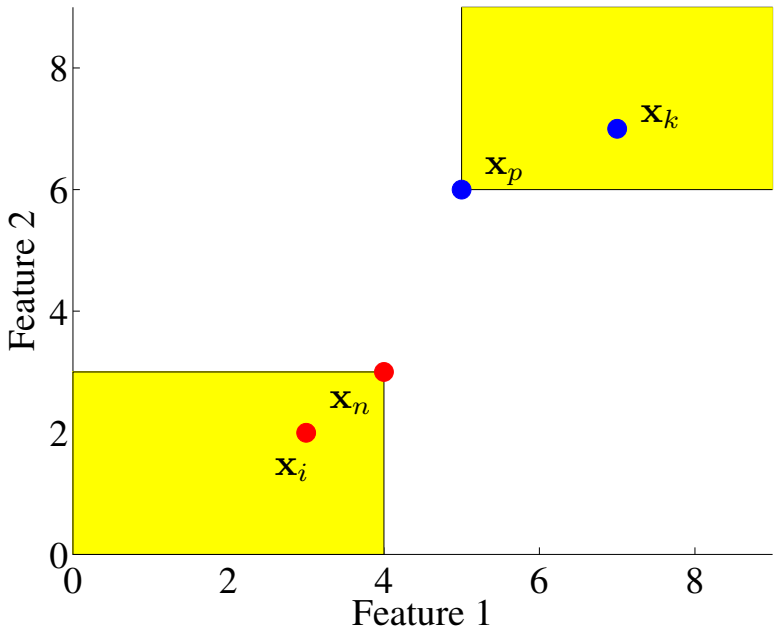
Ранговые шкалы и метки классов

Каждый объект $\mathbf{x} = [\chi_1, \ldots, \chi_j, \ldots, \chi_d]^\mathsf{T}$, описан в ранговых шкалах $\chi_j \in \mathbb{L}_j = \{1 \prec \cdots \prec k_j\}$. На множестве признаков введено отношение частичного порядка.
На множестве $\mathbb{Y} = \{1, 2, 3\}$ меток классов y задано отношение порядка: $1 \prec 2 \prec 3$.

Требуется построить монотонную функцию $\varphi\colon \mathbf{x} \mapsto \hat{y}$

$$\varphi_{opt} = \arg\min_{\varphi} S(\varphi) = \arg\min_{\varphi} \frac{1}{m} \sum_{i \in \mathcal{I}} r\left(y_i, \varphi(\mathbf{x}_i)\right).$$

Отношение доминирования



Без
учета важности признаков

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

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

Объект не доминирует
сам себя ни в одном
из смыслов: $\mathbf{x} \not\succ_n \mathbf{x}$, $\mathbf{x} \not\succ_p \mathbf{x}$.

Отношение доминирования

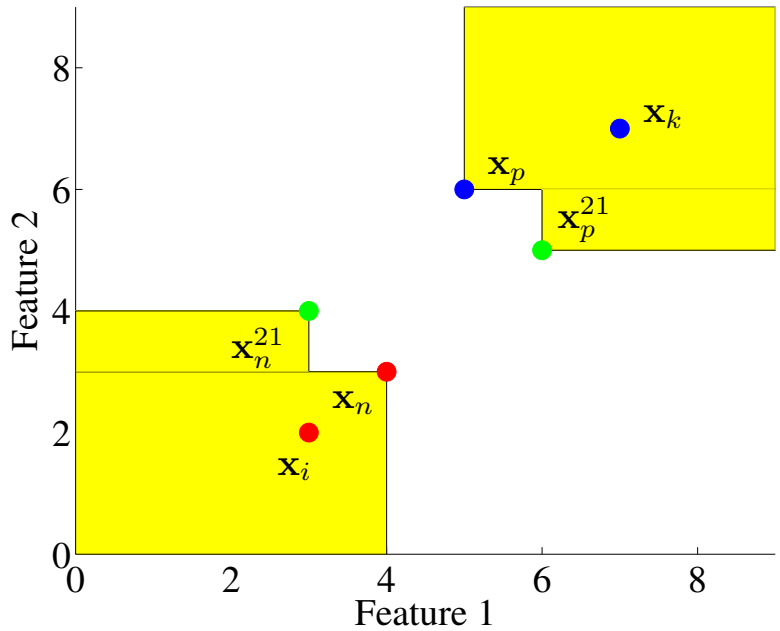
С учетом важности признаков

Пусть признак r важнее, чем признак t.

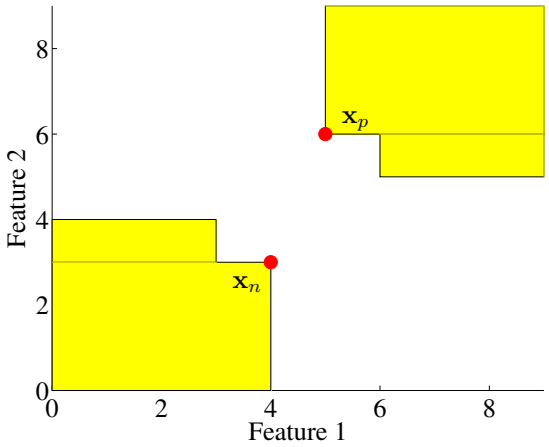
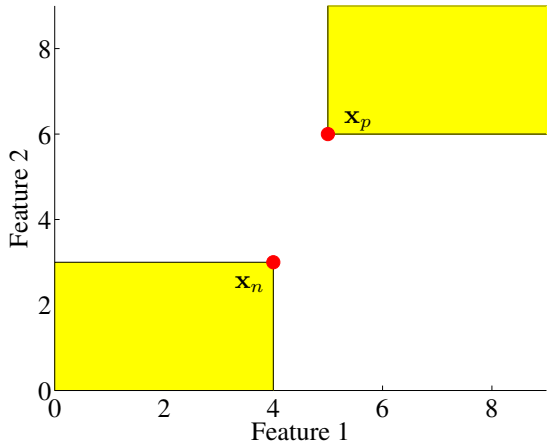
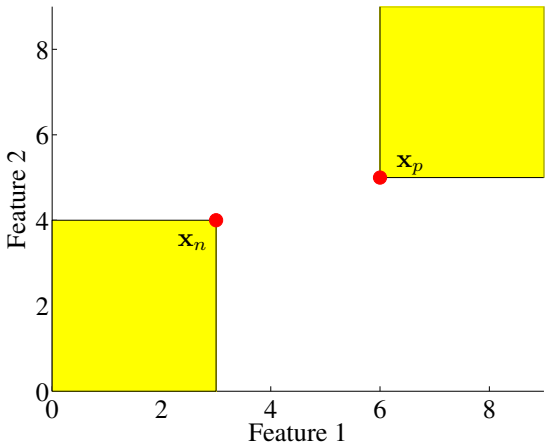
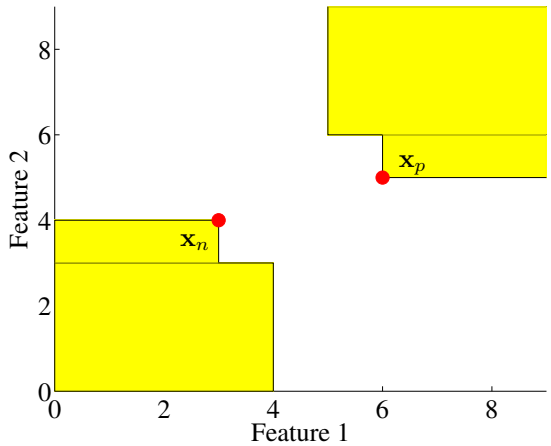
$\mathbf{x}_n \succ_{\tilde{n}} \mathbf{x}_i$, если $\mathbf{x}_n \succ_n \mathbf{x}_i$
или $x_{nr} > x_{nt}$ и $\mathbf{x}_n^{rt} \succ_n \mathbf{x}_i$.

$\mathbf{x}_p \succ_{\tilde{p}} \mathbf{x}_k$, если $\mathbf{x}_p \succ_p \mathbf{x}_k$
или $x_{pr} < x_{pt}$ и $\mathbf{x}_p^{rt} \succ_p \mathbf{x}_k$.

Объект не доминирует
сам себя ни в одном
из смыслов: $\mathbf{x} \not\succ_{\tilde{n}} \mathbf{x}$, $\mathbf{x} \not\succ_{\tilde{p}} \mathbf{x}$.



Области доминирования

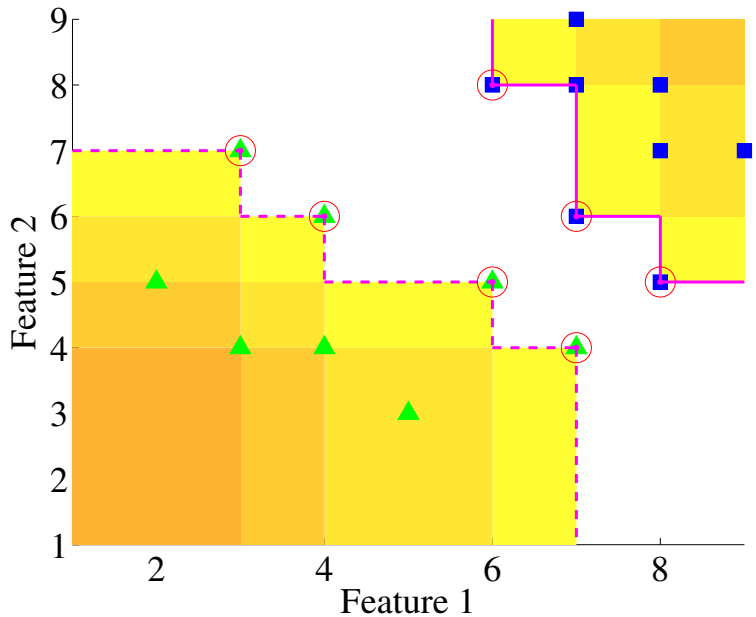
	Признак 1 важнее, чем признак 2	Признак 2 важнее, чем признак 1
$x_{n1} > x_{n2}$, $x_{p1} < x_{p2}$		
$x_{n1} < x_{n2}$, $x_{p1} > x_{p2}$		

Парето-оптимальные фронты

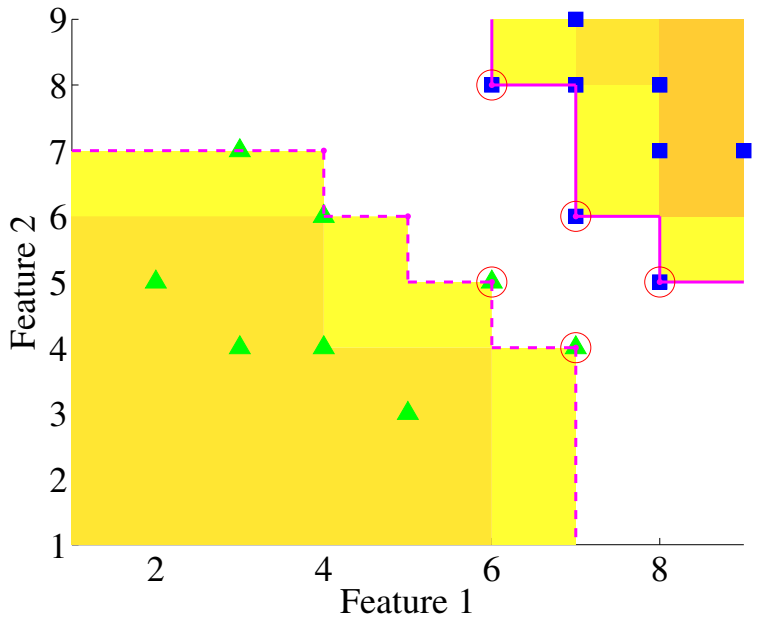
POF_n, POF_p

Множество объектов $\mathbf{x}$, для каждого элемента которого не существует ни одного объекта $\mathbf{x}'$, такого, что

POF_n : $\mathbf{x}' \succ_n \mathbf{x}$ ($\mathbf{x}' \succ_{\tilde{n}} \mathbf{x}$); POF_p : $\mathbf{x}' \succ_p \mathbf{x}$ ($\mathbf{x}' \succ_{\tilde{p}} \mathbf{x}$).



(a)



(b)

Классификация для случая двух классов

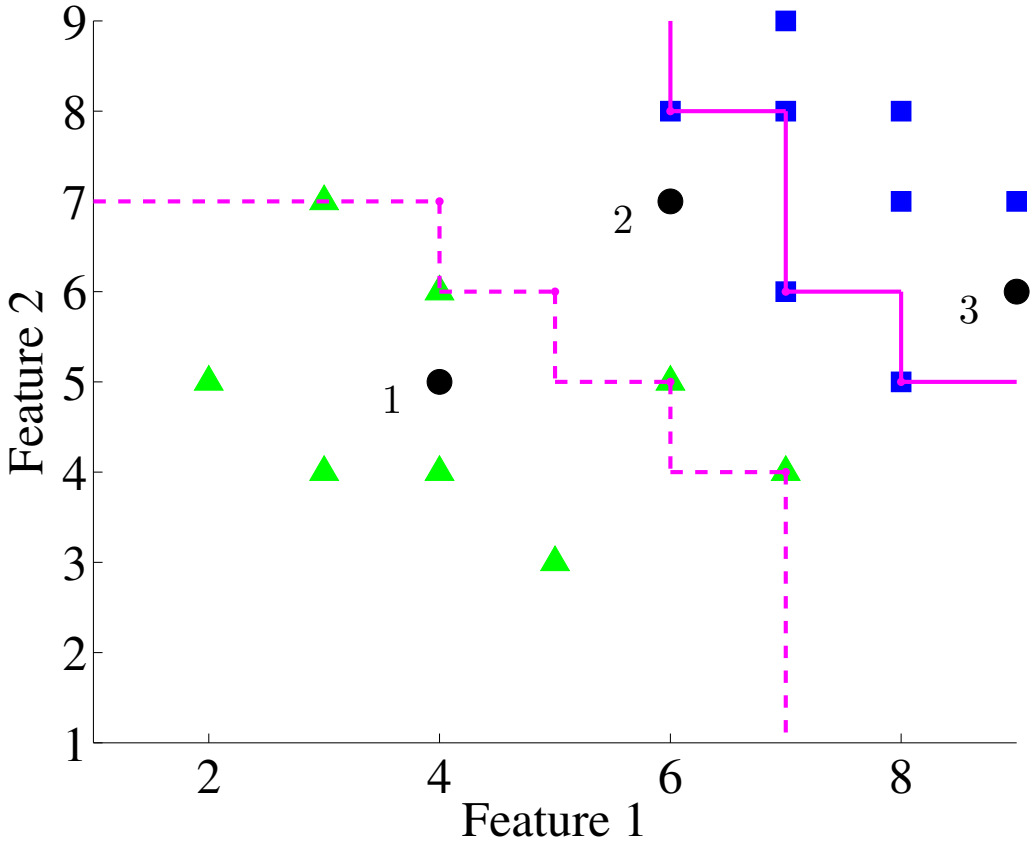
$\mathbf{x}$ — классифицируемый объект
$f(\cdot)$ — классификатор

$$f(\mathbf{x}) = \begin{cases} 0, & \mathbf{x}_n \succ_n \mathbf{x}; \\ 1, & \mathbf{x}_p \succ_p \mathbf{x}; \\ f\left(\underset{\mathbf{x}' \in \overline{\mathrm{POF}}_n \cup \overline{\mathrm{POF}}_p}{\arg\min} \left(\rho(\mathbf{x}, \mathbf{x}')\right)\right), & \text{иначе.} \end{cases}$$

$\overline{\mathrm{POF}}_n, \overline{\mathrm{POF}}_p$ — границы областей доминирования соответствующих Парето-оптимальных фронтов.
ρ — функция расстояния между объектами

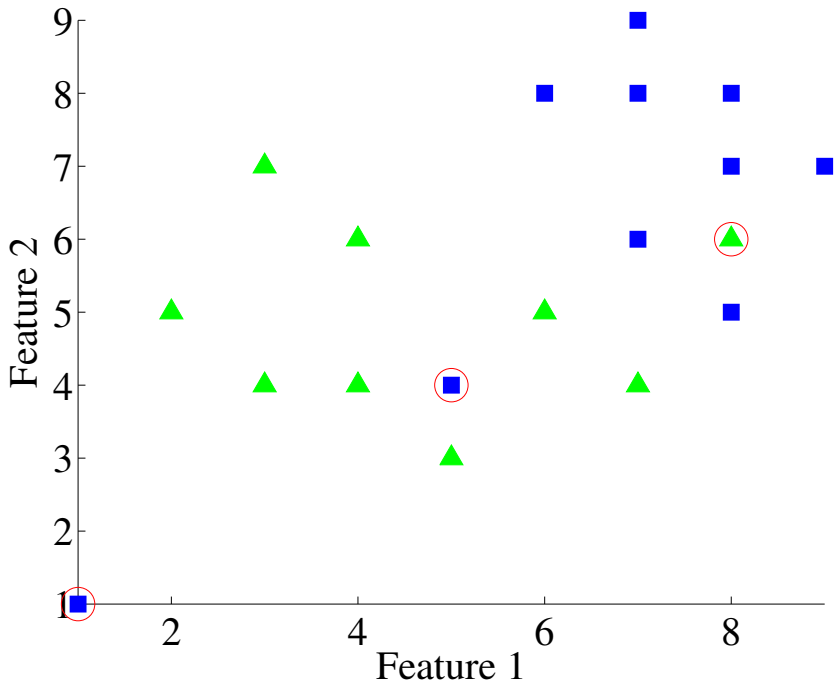
$$\rho(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} r(x_j, x'_j).$$

Пример двухклассовой классификации

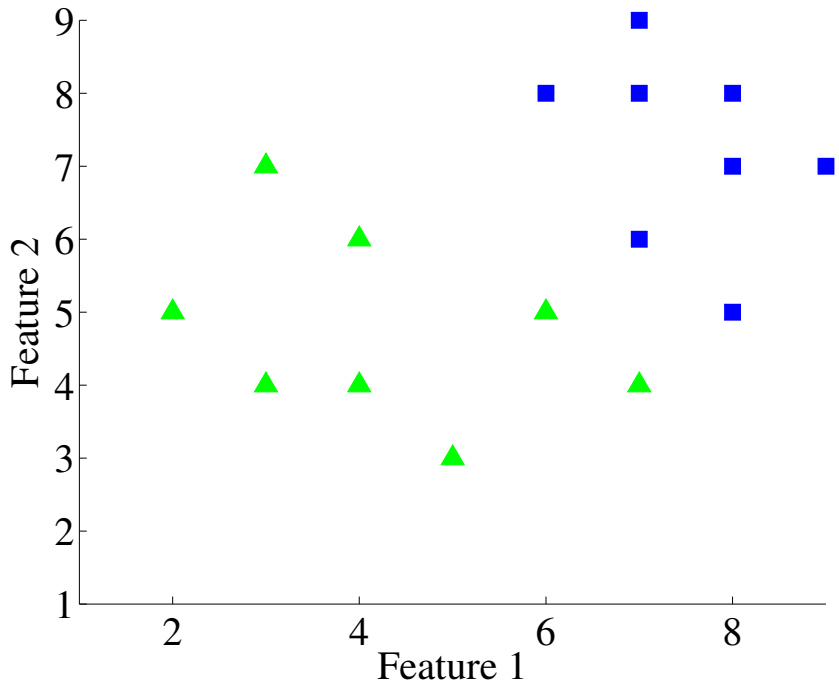


№	Объект **x**	$f(\mathbf{x})$
1	(4,5)	0
2	(6,7)	1
3	(9,6)	1

Приведение выборки к разделимой



(c) С дефектными объектами



(d) Без дефектных объектов

Определение монотонного классификатора

$\{1 \prec \cdots \prec u \prec u+1 \prec \cdots \prec z\} = \mathbb{Z}$ — метки классов

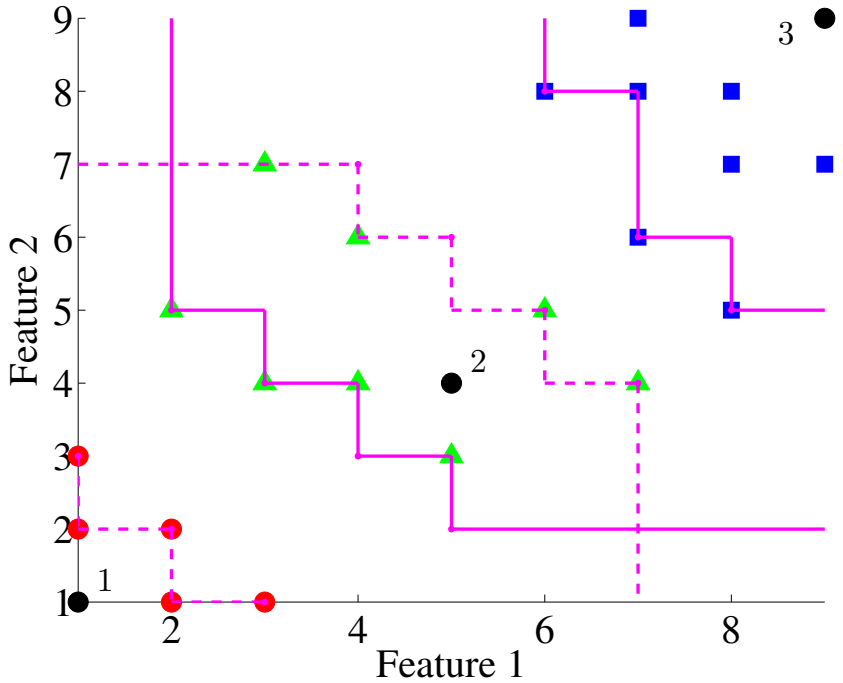
$f_{u,u+1}\colon \mathbf{x} \mapsto \hat{y} \in \{0,1\}$ — двухклассовый классификатор для пары смежных классов

«0» — все классы с метками $y \preceq u$
«1» — все классы с метками $y \succeq u+1$

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

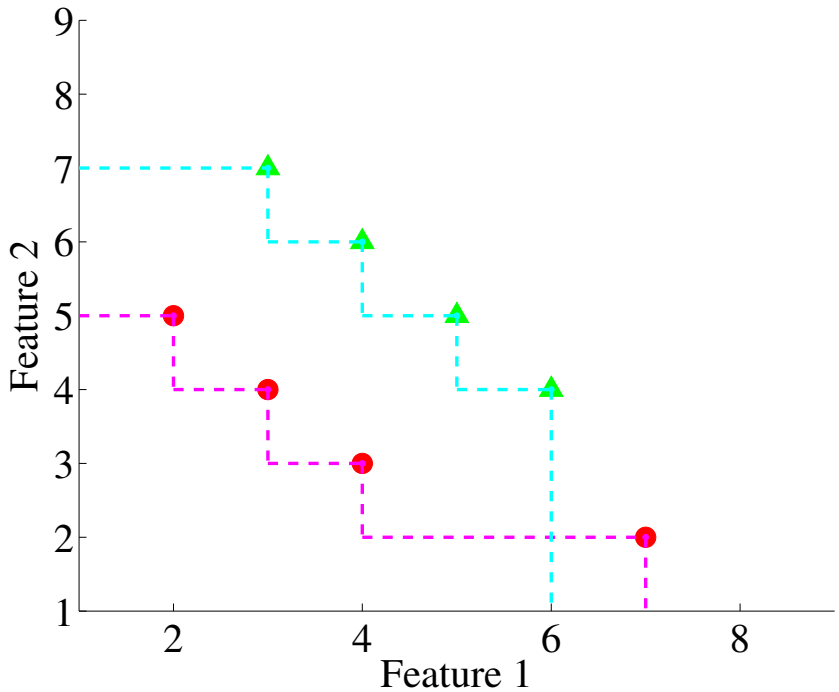
$1,\ 2$	...	$u-1,\ u$	$u,\ u+1$	...	$z-1,\ z$
1	...	1	0	...	0

Пример многоклассовой классификации

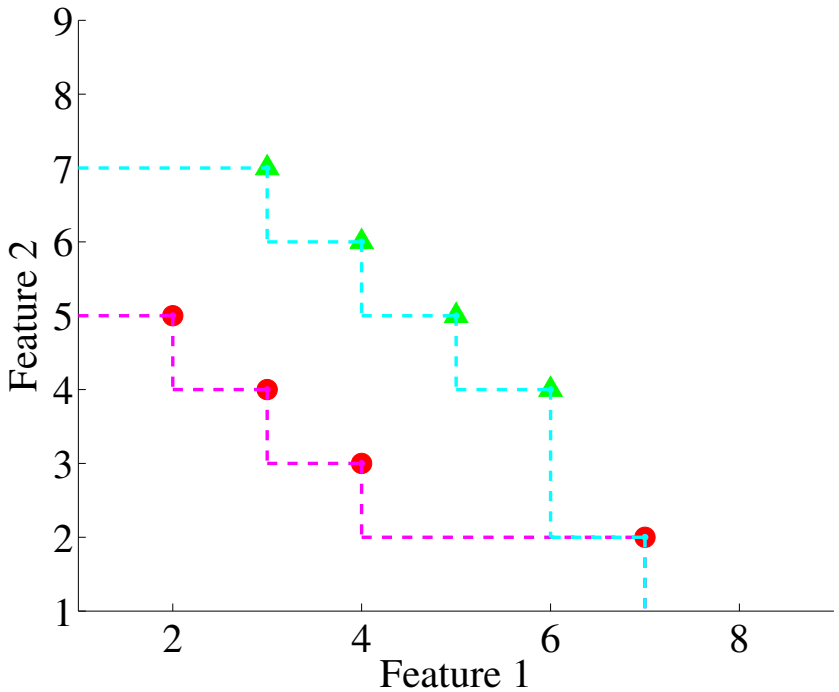


№	Объект $\mathbf{x}$	$f_{12}(\mathbf{x})$	$f_{23}(\mathbf{x})$	$\varphi(\mathbf{x})$
1	(1,1)	0	0	1
2	(5,4)	1	0	2
3	(9,9)	1	1	3

Доопределение фронтов при монотонной классификации



(e) Без доопределения



(f) С доопределением

Общий объект для двух n-фронтов

Допустимые классификаторы

Условие транзитивности

$$\begin{cases} f_{u,u+1}(\mathbf{x}) = 0 \Rightarrow f_{(u+s)(u+1+s)}(\mathbf{x}) = 0 & \text{для всех } s\colon (u+1+s) \leqslant z, \\ f_{u,u+1}(\mathbf{x}) = 1 \Rightarrow f_{(u-s)(u+1-s)}(\mathbf{x}) = 1 & \text{для всех } s\colon (u-s) \geqslant 1. \end{cases}$$

Определение

Классификатор φ *допустимый*, если для всех входящих в него функций $f_{u,u+1}$ соблюдается условие транзитивности.

Теорема

Непересечение фронтов $\mathrm{POF}_n(u)$ и $\mathrm{POF}_p(u+1)$, $u = 1, \ldots, z-1$ влечет выполнение условия транзитивности для любого классифицируемого объекта.

Сравнение алгоритмов

Алгоритм	Средняя ошибка на обучении	LOO	Время построения модели, сек
POF (предлагаемый)	0.22	0.56	2.1
Решающие деревья	0.25	0.69	0.4
Криволинейная регрессия [1]	0.57	0.71	3.6
Конусы [2]	0.29	0.58	1.2
Копулы [3]	0.57	0.61	0.25

[1] 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.

[2] 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.

[3] 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.

Сравнение вычисленных и экспертных категорий

	Вычисленные категории риска		
Class labels	1	2	3
1	Азовская белуга Схизофрагма гортензиевидная Миякея цельнолистная Морская минога Калуга Азовская белуга Сахалинский осетр	Сахалинский осетр Береза Максимовича Гнездовка уссурийская Горал Дальневосточный леопард	Кильдинская треска Нельма подвид нельма Днепровский усач
2	Сибирский осетр западносибирский обский подвид Сахалинский таймень Русская быстрянка Китайский окунь или ауха	Сибирский осетр байкальский подвид Амурский тигр Уссурийский пятнистый олень Украинская минога Сибирский осетр байкальский подвид Волжская сельдь	Волховский сиг Водяной орех чилим Луговик Турчанинова Остролодочник тодомоширский Цетрария степная
3	Куликлопатень Солнцецвет арктический	Белый медведь лаптевская популяция Кизильник киноварнокрасный Маннагеттея Гуммеля Осмунда японская Чистоус японский	Длинноперая палия Световидова Желтозобик Мелколепестник сложноцветный Сердечник пурпурный Калопанакс семилопастный Омфалина гудзонская Малоротая палия

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.

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.
- Предлагаемый алгоритм при использовании всех признаков и информации об их структуре обеспечивает более качественную классификацию по сравнению с другими алгоритмами, применявшихся для решения рассматриваемой задачи.