Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Кузнецов Михаил Павлович

Построение интегральных индикаторов в задачах с порядковыми признаками

511656 - Математические и информационные технологии

Магистерская диссертация

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

Москва
2013

Содержание

Аннотация

Рассмотрена задача построения интегральных индикаторов объектов, описанных в смешанных (линейных, номинальных, порядковых) шкалах. Множеству значений признака ставится в соответствие подмножество евклидового пространства специального вида. В случае порядковой шкалы, множеству значений признака ставится в соответствие конус, описывающий возможные значения этого признака. На основе параметризации этих подмножеств строится новое признаковое описание, и выполняется процедура построения интегральных индикаторов, заключающаяся в поиске ближайших векторов в конусах. Рассматриваются случаи заданного экспертами вектора параметров, значения которого принадлежат линейной и порядковой шкале. Алгоритм используется для категоризации редких видов Красной книги РФ. Производится сравнение результатов алгоритма с двумя альтернативными подходами: статистическим методом и алгоритмом криволинейной регрессии.

Ключевые слова: *интегральный индикатор, ранжирование альтернатив, порядковые шкалы, номинальные шкалы, обучение предпочтений.*

1 Введение

Актуальность темы. Задача построения интегральных индикаторов, или оценок качества объектов, встречается во многих прикладных областях. Особенностью этой задачи является нечисловая природа изучаемых объектов и их оценок. В зависимости от прикладной задачи, качество объектов может измеряться в различных шкалах [1]: номинальных, порядковых, интервальных, абсолютных.

Данная работа фокусируется на рассмотрении порядковых типов шкал. При этом рассматриваются отношения частичного порядка на элементах оценок объектов. Задачи подобного типа носят название задач «обучения предпочтений» (preference learning) [2]. В зависимости от постановки, в таких задачах требуется восстановить линейный или частичных порядок на множестве объектов, располагая обучающей выборкой, состоящей из матрицы «объекты-признаки» и экспертно заданными значениями интегральных индикаторов на объектах обучающей выборки.

Празнаки, описывающие набор объектов, могут также иметь нечисловую природу и измеряться в различных шкалах. В данной работе рассматриваются частично упорядоченные множества заначений признаков. Примером прикладной задачи подобного типа является, например, задача коллаборативной фильтрации [3].

Цель работы. Предложить алгоритм построения интегральных индикаторов показателей качества объектов, основываясь на экспертной информации об объектах обучающей выборки. Объекты описаны признаками, значения которых принадлежат частично упорядоченным множествам.

Методы исследований. При построении алгоритма использовались методы построения монотонных композиций, элементы теории бинарных отношений, методы обучения предпочтений, методы обработки нечисловой информации.

Научная новизна.

- Предложен способ описания порядковой шкалы конусом,
- установлено соответствие между конусом, описывающим частично упорядоченное множество, и матрицей инцидентности графа, соответствующего этому множеству,

- предложен способ корректной бинаризации различных типов шкал,
- на основе бинаризации шкал предложен алгоритм монотонной классификации, или построения интегральных индикаторов, объектов.

Практическая ценность. Разработан программный модуль, который

- на основе обучающей выборки строит алгоритм монотонной классификации объектов,
- оценивает качество предложенного алгоритма,
- иллюстрирует результаты.

Положения, выносимые на защиту:

- Метод корректной бинаризации порядковых шкал,
- алгоритм построения интегральных индикаторов объектов по их порядковому описанию.

Апробация. Результаты работы были использованы для определения статусов редких видов Красной книги РФ [4].

Обзор литературы. Задача построения интегральных индикаторов является одной из задач в области обучения предпочтений (preference learning) [2]. Всего в этой области различают три типа задач. Первый тип задач — это задачи ранжирования меток (label ranking) [5]. Помимо множества объектов, задано множество меток классов. Для каждого объекта необходимо упорядочить метки классов, задав «предпочтения» каждого объекта по классам. Частным случаем такого типа задач служит, например, задача многоклассовой классификации.

Второй тип задач в области обучения предпочтений — задачи монотонной классификации, или порядковой регрессии (instance ranking) [6]. Множество меток классов, или интегральных индикаторов, является частично упорядоченным множеством. Необходимо поставить каждому объекту в соответствие элемент этого множества, основываясь на признаковом описании объекта и обучающей информации.

Задачи третьего типа являются задачами ранжирования объектов [7]. Этот тип отличается от предыдущего тем, что не существует отдельного множества значений

классов. Ответом алгоритма на наборе объектов должна служить функция, ранжирующая эти объекты. Например, это может быть граф частичного порядка на множестве объектов.

Рассматривая в данной работе задача близка к задачам второго типа, то есть к задачам монотонной классификации. Задача монотонной классификации, или порядковой регрессии, имеет широкое применение в сфере информационного поиска [8], и потому является тщательно изученной. Принципиальное отличие данной задачи от общепринятой постановки состоит в том, что значения признаков, описывающих объекты, принадлежат частично упорядоченным множествам.

Принадлежность значений признаков частично упорядоченным множествам является одной из задач обработки нечисловой информации [9]. В задачах машинного обучения развивается направление анализа формальных понятий [10], согласно которому, объекты описываются нечисловыми атрибутами. Признаки подобного типа описываются, в том числе, т.н. решетками Галуа.

В работе рассматриваются частично упорядоченные множества значений признаков. Множества подобного типа характеризуются тем, что на них введено бинарное отношение, обладающее свойствами рефлексивности, антисимметричности и транзитивности [11]. Работа опирается на исследования в области аггрегирования бинарных отношений и частичных порядков [12], активно применяющихся, например, в задачах голосования.

Частично упорядоченные множества в работе описываются выпуклыми конусами [13]. Производится процедура бинаризации частичных порядков, связанная с тем, что любая точка внутри конуса представляется выпуклой комбинацией образующих конуса. Матрица, соответствующая бинаризованному признаку, сопоставляется с матрицей инцидентности графа, соответствующему частичному порядку. Новое признаковое описание используется для построения монотонного классификатора, основанного на линейной модели.

Предлагается итеративный алгоритм оценки параметров модели. Результаты работы алгоритма сравниваются с двумя альтернативными подходами. Первый подход, статистический, использует копулы для анализа совместного распределеня порядковых признаков [14]. Второй подход состоит в обобщении метода линейной регрессии на случай порядковых признаков [18].

2 Постановка задачи построения интегральных индикаторов

В этом разделе сформулируем постановку задачи построения интегральных индикаторов и обобщим эту постановку на случай порядковых шкал.

2.1 Модель построения интегральных индикаторов

Дана выборка наблюдений $\{\mathbf{X}, \mathbf{y}_0\}$. $\mathbf{X}$ является матрицей плана размера $m \times n$ и представляется набором своих столбцов:

$$\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n],$$

где в случае линейных шкал $\boldsymbol{\chi}_j \in \mathbb{R}^m$. Случай более общего типа шкал будет рассмотрен позднее.

Вектор $\mathbf{y}_0$ является экспертной оценкой множества интегральных индикаторов объектов. Задача построения интегральных индикаторов заключается в поиске модели $f(\mathbf{X}, \mathbf{w})$, такой, что

$$\mathbf{y}_1 = f([\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n], \mathbf{w}).$$

Оптимальные параметры $\hat{w}$ должны минимизировать функцию ошибки

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} S(f([\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n], \mathbf{w}), \mathbf{y}_0).$$

Сложность задачи заключается в том, что множества значений признаков $\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n$, а также множество значений вектора интегральных индикаторов $\mathbf{y}$ не являются Евклидовыми пространствами. В частности будем говорить, что множество значений признака принадлежит порядковой шкале, если оно частично упорядочено.

2.2 Порядковые шкалы: частично упорядоченные множества

Будем считать, что элементы x_{ij} матрицы описаний $\mathbf{X}$ принадлежат частично упорядоченным множествам $\mathbb{L}_j$ мощности L_j. Для упрощения обозначений, на время опустим индекс j и будем работать с одним частично упорядоченным множеством $\mathbb{L}$ мощности L:

$$\mathbb{L} = \{l_1, ..., l_L\}, \quad |\mathbb{L}| = L, \tag{2.1}$$

с определенной на $\mathbb{L}$ бинарной операцией $\succcurlyeq$. Согласно определению частично упорядоченного множества, бинарная операция $\succcurlyeq$ обладает тремя свойствами:

- рефлексивность, $(l \succcurlyeq l)\ \forall l \in \mathbb{L}$;
- транзитивность, $(l_{k_1} \succcurlyeq l_{k_2}) \wedge (l_{k_2} \succcurlyeq l_{k_3}) \Rightarrow (l_{k_1} \succcurlyeq l_{k_3})\ \forall l_{k_1}, l_{k_2}, l_{k_3} \in \mathbb{L}$;
- антисимметричность, $(l_{k_1} \succcurlyeq l_{k_2}) \wedge (l_{k_2} \succcurlyeq l_{k_1}) \Rightarrow (l_{k_1} = l_{k_2})\ \forall l_{k_1}, l_{k_2} \in \mathbb{L}$.

Каждому частично упорядоченному множеству $\mathbb{L}$ соответствует ориентированный граф, вершинами которого являются элементы $l_1, ..., l_K$ множества $\mathbb{L}$. Между вершинами l_{k_1}, l_{k_2} графа существует ребро, исходящее из l_{k_1}, если выполнено $l_{k_1} \succcurlyeq l_{k_2}$. Такой граф описывается матрицей инцидентности $\mathbf{L}$ размера $L \times L$:

$$\mathbf{L}(k_1, k_2) = \begin{cases} 1, & \text{если } l_{k_1} \succeq l_{k_2}, \\ 0, & \text{иначе.} \end{cases} \tag{2.2}$$

Матрица $\mathbf{L}$ будет играть важнейшую роль в дальнейшем описании порядковых признаков.

3 Описание порядковой шкалы конусом

В этой секции поставим в соответствие частично упорядоченному множеству конус в Евклидовом пространстве. Параметризовав этот конус, поставим задачу построения интегральных индикаторов.

3.1 Соответствие полиэдрального конуса порядковому признаку

Пусть матрица описаний $\mathbf{X}$ состоит из столбцов $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$, где $\boldsymbol{\chi}_j \in \mathbb{L}_l^m$. Здесь m — декартова степень множества $\mathbb{L}_j$, равная количеству объектов в выборке. Столбцу $\boldsymbol{\chi}$, значения которого принадлежат частично упорядоченному множеству $\mathbb{L}$, поставим в соответствие подмножество Евклидового пространства $\chi \in \mathbb{R}_+^m$:

$$\chi = \{\mathbf{x} | \quad \mathbf{x} \in \mathbb{R}_+^m, \quad \boldsymbol{\chi}_{k_1} \succcurlyeq \boldsymbol{\chi}_{k_2} \Rightarrow x_{k_1} \geqslant x_{k_2}\}. \tag{3.1}$$

Отметим, что множество $\chi \subset \mathbb{R}_+^m$ и вектор $\boldsymbol{\chi} \in \mathbb{L}_l^m$ являются разными сущностями и обозначаются одинаково с целью избежать избыточных символов.

Теорема 3.1. *Множество $\chi \subset \mathbb{R}_+^m$, определенно системой неравенств (3.1), является полиэдральным конусом в Евклидовом пространстве $\mathbb{R}^m$ размерности, не превышающей L.*

Доказательство.

Во-первых, отметим, что определение (3.1) введенно корректно в силу условий транзитивности, рефлексивности и антисимметричности, налагаемых на бинарное отношение $\succcurlyeq$.

То, что множество χ является полиэдральным конусом, следует из определения полиэдрального конуса, задаваемого системами неравенств.

Количество различных значений компонент вектора $\boldsymbol{\chi}$ не может превышать мощности L частично упорядоченного множества $\mathbb{L}$. Значит, вектор $\boldsymbol{\chi}$ состоит из повторяющихся элементов (в случае $m > L$), и вследствие антисимметричности в системе неравенств (3.1) появляются равенства $\boldsymbol{\chi}_{k_1} = \boldsymbol{\chi}_{k_2} \Rightarrow x_{k_1} = x_{k_2}$, разбивая множество значений признака $\boldsymbol{\chi}$ на множества одинаковых компонент. Количество таких множеств не превышает L и совпадает с размерностью конуса χ. ■

Представление конуса линейной комбинацией образующих. Для того, чтобы решать задачу построения интегральных индикаторов, параметризуем конусы χ_j, соответствующие столбцам $\boldsymbol{\chi}_j$ матрицы $\mathbf{X}$.

Теорема 3.2. *В случае, когда собственная размерность конуса χ равна L, вектор $\mathbf{x}$, принадлежащий конусу χ, соответствующий признаку $\boldsymbol{\chi}$, представляется неотрицательной комбинацией образующих конуса*

$$\mathbf{x} = w\sum_{k=1}^{L}\lambda_k\boldsymbol{\zeta}_k, \quad w \geqslant 0, \quad \sum_{k=1}^{L}\lambda_k = 1, \quad \lambda_k \geqslant 0,$$

где $\boldsymbol{\zeta}_k$ — образующая конуса χ,

$$\boldsymbol{\zeta}_k(i) = \begin{cases} 1, & \text{если } \boldsymbol{\chi}_i \succcurlyeq l_k, \\ 0, & \text{иначе,} \end{cases}$$

причем такое разложение единственно.

Доказательство.

Отметим, что поскольку собственная размерность конуса χ равна L, конус определяется в точностью L образующими, то есть формула разложения определена корректно. Само разложение по образующим следует из теоремы Каратеодори о выпуклой оболочке. Поскольку размерность конуса L не превышает размерности пространства m, конус χ является симплициальным, и такое разложение единственно. ■

Отметим, что в случае размерности конуса χ меньшей L теорема является выполненной, однако количество образующих в этом случае совпадает с размерностью конуса.

Рассмотрим пример бинаризации матрицы описаний $\mathbf{X}$.

Пример 3.1. Пусть

$$\mathbf{X} = \begin{pmatrix} l_{11} & l_{23} \\ l_{12} & l_{22} \\ l_{13} & l_{21} \end{pmatrix}, \quad \mathbb{L}_1 = \{l_{11} \succeq l_{12} \succeq l_{13}\}, \mathbb{L}_2 = \{l_{21} \succeq l_{22} \succeq l_{23}\}.$$

Тогда матрица $\mathbf{X}$ записывается следующим образом:

$$\mathbf{X} = [\chi_1, \chi_2],$$

$$\chi_1 = \lambda_{11}\boldsymbol{\zeta}_{11} + \lambda_{12}\boldsymbol{\zeta}_{12} + \lambda_{13}\boldsymbol{\zeta}_{13} = \lambda_{11}\begin{pmatrix}1\\1\\1\end{pmatrix} + \lambda_{12}\begin{pmatrix}1\\1\\0\end{pmatrix} + \lambda_{13}\begin{pmatrix}1\\0\\0\end{pmatrix},$$

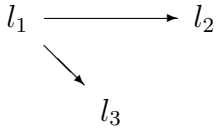
$$\chi_2 = \lambda_{21}\boldsymbol{\zeta}_{21} + \lambda_{22}\boldsymbol{\zeta}_{22} + \lambda_{23}\boldsymbol{\zeta}_{23} = \lambda_{21}\begin{pmatrix}1\\1\\1\end{pmatrix} + \lambda_{22}\begin{pmatrix}0\\1\\1\end{pmatrix} + \lambda_{23}\begin{pmatrix}0\\0\\1\end{pmatrix},$$

при выполнении условий

$$\sum_{k=1}^{L_j} \lambda_{jk} = 1, \quad \lambda_{jk} \geqslant 0. \tag{3.2}$$

Согласно определению образующих $\boldsymbol{\zeta}_k$, элементы этих образующих являются, по сути, элементами матрицы инцидентами графа, соответствующего частичному порядку. Проиллюстрируем это на примере.

Пример 3.2.



(3.3)

Рассмотрим пример частично упорядоченного множества $\mathbb{L}$, задаваемого графом (3.3). Это множество состоит из трех элементов l_1, l_2, l_3, для которого выполня-

ются условия $l_1 \succeq l_2$, $l_1 \succeq l_3$. Согласно определению (2.2), матрица $\mathbf{L}$, соответствующая графу (3.3),

$$\mathbf{L} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Строки этой матрицы используются при формировании матрицы $\mathbf{Z}$ для заданного набора объектов.

Из теоремы 3.2 и примера 3.2 видно, что для бинаризации частичного порядка и составления образующих конуса достаточно воспользоваться матрицей смежности графа, соответствующего частично упорядоченному множеству.

3.2 Нахождение оптимальных параметров модели в случае порядковых признаков

Задача построения интегральных индикаторов заключается в поиске модели $f(\mathbf{X}, \mathbf{w})$, такой, что

$$\mathbf{y}_1 = f([\chi_1, ..., \chi_n], \mathbf{w}).$$

Оптимальные параметры $\hat{w}$ должны минимизировать функцию ошибки

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} S(f([\chi_1, ..., \chi_n], \mathbf{w}), \mathbf{y}_0).$$

Будем рассматривать линейные модели, то есть вектор $\mathbf{y}_1$ представляется линейной комбинацией

$$\mathbf{y}_1 = w_1\chi_1 + ... + w_n\chi_n.$$

Множества χ_j являются подмножествами $\mathbb{R}^m$, однако благодаря параметризации, задаваемой теоремой 3.2, линейная модель запишется в следующем виде:

$$\mathbf{y}_1 = w_1\boldsymbol{\lambda}_1\boldsymbol{Z}_1 + ... + w_n\boldsymbol{\lambda}_n\boldsymbol{Z}_n, \tag{3.4}$$

где векторы $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ являются векторами элементов выпуклых комбинаций конусов χ_j,

$$w_j \in \mathbb{R}, \qquad \boldsymbol{\lambda}_j \in \Lambda = \{\boldsymbol{\lambda} | \boldsymbol{\lambda} \geqslant \mathbf{0}, \|\boldsymbol{\lambda}\|_1 = 1\},$$

а столбцы матриц $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ являются образующими конусов $\chi_1, ..., \chi_n$, определяемыми теоремой 3.2.

Оптимальные параметры $\mathbf{w}$ должны минимизировать функцию ошибки:

$$S(\mathbf{f}([\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n], \mathbf{w}), \mathbf{y}_0) \to \min_{\substack{\mathbf{w} \in \mathbb{R}^n; \\ \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0 \in \Lambda}}.$$

В линейном случае (3.4), а также в случае квадратичной функции ошибки,

$$S(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|,$$

задача поиска оптимальных параметров запишется следующим образом:

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}_1, ..., \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0) = \arg\min_{\substack{\mathbf{w} \in \mathbb{R}^n; \\ \boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0 \in \Lambda}} \|\boldsymbol{\lambda}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \boldsymbol{\lambda}_j \boldsymbol{Z}_j\|. \tag{3.5}$$

Отметим, что в данном случае экспертная оценка $\mathbf{y}_0$ выставлена в ранговой шкале, и ее параметризация также дается теоремой 3.2 с матрицей образующих $\boldsymbol{Z}_0$ и соответствующим вектором выпуклой комбинации $\boldsymbol{\lambda}_0$.

Алгоритм вычисления оптимальных параметров. Отметим, что оптимальные параметры, соответствующие минимизации (3.5), не вычисляются явным образом. Предлагается итеративный алгоритм. На каждом четном шаге при фиксированных параметрах выпуклой комбинации $\hat{\boldsymbol{\lambda}}_1, ..., \hat{\boldsymbol{\lambda}}_n, \hat{\boldsymbol{\lambda}}_0$ вычисляется вектор параметров модели $\mathbf{w}$, затем при фиксированном $\hat{\mathbf{w}}$ вычислются $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n, \boldsymbol{\lambda}_0$.

Принимается начальное приближение алгоритма: $\hat{\lambda}_{jk} = 1/L_j$, то есть образующие конусов имеют одинаковый вес, и рассматриваются центральные точки всех конусов.

На каждом следующем шаге производится процедура последовательной минимизации функции ошибки:

1. $\hat{\mathbf{w}} := \arg\min_{\mathbf{w} \in \mathbb{R}} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|,$
2. $\hat{\boldsymbol{\lambda}}_j = \arg\min_{\boldsymbol{\lambda}_j \in \Lambda} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} \hat{w}_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|$, для всех $j = 1, ..., n$.

Выполнена теорема о то скорости нахождения алгоритмом оптимального решения.

Теорема 3.3. *В случае линейной модели 3.4, итеративный алгоритм находит оптимальное решение не более чем за Θ итераций, где*

$$\Theta = L_0 - 1 + L_1 - 1 + ... + L_n - 1 + n = \sum_{j=0}^{n} L_j - 1 \quad -$$

собственная размерность пространства параметров.

Доказательство.

Идея доказательства заключается в следующем: на каждой итерации алгоритм находит в соответствующем конусе вектор, ближайший к линейной комбинации остальных. В самом трудном случае, когда конусы не пересекаются, оптимальные векторы лежат на образующих или гранях конусов. На каждом шаге алгоритм выбирает в каждом конусе соответствующую грань, не ухудшая решение. Всего количество граней всех конусов совпадает с собственной размерностью пространства признаков Θ, то есть и количество шагов алгоритма не превосходит Θ. ■

Множество параметров модели. Рассмотрим более детально множество параметров $\mathbf{w}$ модели f. Согласно теореме 3.2, вектор $\mathbf{w} \in \mathbb{R}_+$ является неотрицательным, однако минимизация ошибки 3.5 проводится по всему множеству $\mathbb{R}$. Такое обобщение позволяет учитывать также отрицательно монотонную зависимость интегральных индикаторов $\mathbf{y}_1$ от признаков. Если предполагается исключительно неотрицательная зависимость, то шаг 1 итеративного алгоритма перепишется следующим образом:

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}\in\mathbb{R}_+} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|.$$

На множестве параметров модели могут быть заданы экспертные предпочтения. Например, эксперт может задать линейных порядок на элементах множества параметров, указав важность признаков, например, так:

$$w_1 \geqslant ... \geqslant w_n \geqslant 0.$$

В этом случае, минимизация параметров модели w может проводиться по экспертно заданному множеству $\mathcal{R} = \{w_1 \geqslant ... \geqslant w_n \geqslant 0\}$, и шаг 1 алгоритма перепишется следующим образом:

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}\in\mathcal{R}} \|\hat{\boldsymbol{\lambda}}_0 \boldsymbol{Z}_0 - \sum_{j=1}^{n} w_j \hat{\boldsymbol{\lambda}}_j \boldsymbol{Z}_j\|.$$

В более общем случае, множество $\mathcal{R}$ может быть произвольным экспертно заданным множество $\mathcal{R} \in \mathbb{R}^n$.

3.3 Непараметрическое описание конусов, соответствующих порядковым признакам

Рассмотрим метод построения интегрального индикатора $\mathbf{y}_1$ по матрице описаний $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$ и экспертной оценке $\mathbf{y}_0$, не вводя множество дополнительных

параметров [14].

В нашем случае, столбцы матрицы $\mathbf{X}$ являются конусами $\chi_1, ..., \chi_n$. Любой полиэдральный конус определяется системой линейных неравенств

$$\chi_j = \{\mathbf{x}_j | A_j \mathbf{x}_j \leqslant 0\}, j = 1, ..., m.$$

В общем случае, множество χ_j является многогранником, определяемым матрицей A_j. Например, в случае линейного порядка на элементах вектора $\boldsymbol{\chi}_j$, $x_{1j} \geqslant x_{2j} \geqslant ... \geqslant x_{mj} \geqslant 0$, матрица

$$A_j = \begin{pmatrix} -1 & 1 & 0 & ... & 0 \\ 0 & -1 & 1 & ... & 0 \\ 0 & 0 & 0 & ... & -1 \end{pmatrix},$$

эта матрица определяет выпуклый многогранный конус χ_j.

Принимается линейная модель построения интегрального индикатора. Это означает, что вектор $\mathbf{y}_1$ представляется в виде суммы:

$$\mathbf{y}_1 = \mathbf{x}_1 + ... + \mathbf{x}_m, \mathbf{x}_1 \in \chi_1, ..., \mathbf{x}_m \in \chi_m.$$

Для дальнейшего рассмотрения напомним понятие суммы Минковского двух множеств. Суммой Минковского двух подмножеств L_1 и L_2 линейного пространства L называется множество L', состоящее из всевозможных сумм векторов из L_1 и L_2. Из этих двух определений логично вытекает определение области значений линейной модели. Областью значений для линейной модели назовем сумму Минковского

$$\chi = \chi_1 + ... + \chi_m.$$

Заметим, что введенное таким образом определение линейной модели с ранговым описанием объектов обобщает линейную модель в ее стандартном понимании:

$$\mathbf{y} = \sum_{j=1}^{m} \mathbf{X}_j w_j,$$

поскольку каждое слагаемое $\mathbf{X}_j w_j$ для всех w_j является подмножеством конуса χ_j. Следовательно, в случае рангового описания исчезают веса признаков.

Таким образом, для построения интегрального индикатора требуется построить сумму Минковского выпуклых многогранников. Для ее построения воспользуемся методом точного построения суммы Минковского двух многогранников, заданных системами линейных неравенств [15].

Итак, пусть дано два выпуклых многогранника, задающихся системами неравенств:

$$\chi_1 = \{\mathbf{x}_1 | A_1\mathbf{x}_1 \leqslant \mathbf{b}_1\}, \chi_2 = \{\mathbf{x}_2 | A_2\mathbf{x}_2 \leqslant \mathbf{b}_2\}.$$

Их суммой Минковского будет являться вектор $\mathbf{x}$, являющийся решением системы:

$$\begin{cases} \mathbf{x} - \mathbf{x}_1 - \mathbf{x}_2 = 0, \\ A_1\mathbf{x}_1 \leqslant \mathbf{b}_1, \\ A_2\mathbf{x}_2 \leqslant \mathbf{b}_2, \end{cases}$$

которая заменой переменной $\mathbf{x}_1 = \mathbf{x} - \mathbf{x}_2$ преобразуется в систему:

$$\begin{cases} A_1\mathbf{x} - A_2\mathbf{x}_2 \leqslant \mathbf{b}_1, \\ A_2\mathbf{x}_2 \leqslant \mathbf{b}_2. \end{cases}$$

Справедливо следующее утверждение: вектор $\mathbf{x} \in \chi$ тогда и только тогда, когда найдется вектор $\mathbf{x}_2$, удовлетворяющий системе 3.3.

Таким образом, задача поиска вектора $\mathbf{x}$ сводится к решению системы линейных неравенств:

$$C\mathbf{x}_2 \leqslant \mathbf{d}, C = \begin{pmatrix} -A_1 \\ A_2 \end{pmatrix}, \mathbf{d} = \begin{pmatrix} \mathbf{b}_1 - A_1\mathbf{x} \\ \mathbf{b}_2 \end{pmatrix}. (2)$$

Для решения этой системы используем вариант леммы Минковского-Фаркаша [16], который формулируется следующим образом.

Пусть A и b — матрица и вектор. Разрешимость системы линейных неравенств $Ax \leqslant b$ эквивалентна тому, что $yb \geqslant 0$ для любого вектора-строки $y \geqslant 0$ со свойством $yA = 0$. В нашем случае, эта лемма записывается так:

$$\exists \mathbf{x}_2 : C\mathbf{x}_2 \leqslant \mathbf{d} \Leftrightarrow \forall \mathbf{z} : C^T\mathbf{z} = 0, \mathbf{z} \geqslant 0 \rightarrow (\mathbf{d}, \mathbf{z}) \geqslant 0.$$

Пусть V является фундаментальной системой решений для этого случая, причем $V = \binom{V_1}{V_2}$, где V_i — ФСР, соответствующая матрице A_i. Тогда условие $(\mathbf{d}, \mathbf{z}) \geqslant 0$ перепишется в виде:

$$V_1^T(\mathbf{b}_1 - A_1\mathbf{x}) + V_2^T\mathbf{b}_2 \geqslant 0.$$

Таким образом, принимая $A = V_1^T A_1$, $\mathbf{b} = V_1^T\mathbf{b}_1 + V_2^T\mathbf{b}_2$, мы получим параметры $A, \mathbf{b}$ системы неравенств, описывающей сумму Минковского $\chi_1 + \chi_2$.

Заметим, что отдельной трудностью является нахождение фундаментальной системы решений V, каждый столбец которой должен быть $\geqslant 0$. Алгоритмы отыскания такой системы описан в [17].

Проекцией точки x на множество D называется вектор $P(x) \in D$, удовлетворяющий условию:

$$P_D(x) = \arg\min_{y \in D} \|x - y\|.$$

Построив область значения модели χ, определим уточненный интегральный индикатор $\mathbf{y}_1$ как проекцию экспертной оценки $\mathbf{y}_0$ на χ:

$$\mathbf{y}_1 = P_\chi(\mathbf{y}_0).$$

Отметим, что эта проекция является единственной в силу выпуклости множества χ.

Рассмотрим небольшой пример. Пусть матрица

$$X = \begin{pmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix}.$$

Тогда для нее матрицы

$$A_1 = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}, A_2 = \begin{pmatrix} 0 & 1 & -1 \\ 1 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix}.$$

Используя матрицу $C = \begin{pmatrix} 0 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$, получаем

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

и неравенство для суммы: $\chi = \{\mathbf{x} | A\mathbf{x} \leqslant 0\}$.

Таким образом, область значений линейной модели записывается в виде:

$$\chi = \begin{cases} \mathbf{x}_1 \geqslant 0, \\ \mathbf{x}_2 \geqslant 0, \\ \mathbf{x}_3 \geqslant 0, \\ \mathbf{x}_1 + \mathbf{x}_3 \geqslant \mathbf{x}_2. \end{cases}$$

Если, например, экспертная оценка $\mathbf{y}_0 = (1, 5, 2)$, то ее проекция на область значения χ: $\mathbf{y}_1 = (\frac{5}{3}, \frac{16}{3}, \frac{11}{3})$.

4 Альтернативные методы

В этом разделе рассмотрим альтернативные подходы к решению задачи построения интегральных индикаторов объектов, описанных в порядковых шкалах. Рассмотрим статистический подход, основанный на совместном распределении критериев [18], а также алгоритм криволинейной регрессии [19]. В этой секции будем рассматривать только линейно упорядоченные множества значений признаков.

4.1 Построение интегрального индикатора в порядковых шкалах с использованием копул

Предлагается найти отображение $a : X \to Y$, минимизирующее функционал среднего риска. Минимум среднего риска достигается алгоритмом

$$a(x) = \arg\max_{y \in Y} P(y|\mathbf{x}_i),$$

где $P(y|\mathbf{x}_i)$ — апостериорная вероятность класса y для объекта $\mathbf{x}$. Эта вероятность является условной по $\mathbf{x}$. Для оценки апостериорной вероятности $P(y|\mathbf{x}_i)$ будем использовать копулы.

Свойства копул, используемые для оценки условной вероятности.

Определение 4.1. *Функция $C : [0, 1]^d \to [0, 1]$ называется копулой размерности d, если выполняются следующие условия:*

$$C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0,$$

$$C(1, \ldots, 1, u, 1, \ldots, 1) = u,$$

$$B = \prod_{i=1}^{d} [a_i, b_i] \subseteq [0, 1]^d : \quad \int_B dC(u) \geqslant 0.$$

Выполнение этих свойств означает, что функция C является функцией распределения многомерной случайной величины $[u_1, \ldots, u_d]$, такой, что одномерное распределение каждого из u_i равномерно на интервале $[0, 1]$.

Важным фактом, позволяющим применять копулы для построения регрессионных моделей, является следующая теорема.

Теорема 4.1. *Многомерная функция распределения случайной величины:*

$$H(x_1, \ldots, x_d) = P[X_1 \leqslant x_1, \ldots, X_d \leqslant x_d]$$

случайного вектора $(X_1, \ldots, X_d)$ *с одномерными функциями распределения*

$$F_i(x) = P[X_i \leqslant x_i]$$

может быть записана в виде:

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

Таким образом, для оценивания совместного распределения H случайных величин $X_1, \ldots, X_d$ достаточно оценить их одномерные распределения $F_i(x_i)$ и функцию копулы, связывающую эти случайные величины.

Следующая теорема утверждает, что функция копулы не изменяется при действии на случайные величины любых монотонных преобразований.

Теорема 4.2. *Пусть* X, Y *— две случайные величины с совместной функцией распределения* $H(x, y)$. *Пусть также* φ, ψ*— две монотонных функции, преобразующие случайные величины* X *и* Y *в*

$$Z = \varphi(X), \quad T = \psi(Y)$$

с совместной функцией распределения $H'(Z, T)$. *Тогда копула, связывающая случайные величины* Z *и* T*:*

$$C'\big(F'(z), G'(t)\big) = H'(z, t) = C(F'(z), G'(t)),$$

то есть,

$$C' = C.$$

Таким образом, чтобы оценить функцию копулы, описывающую связь между случайными величинами $X_1, \ldots, X_d$, достаточно знать только ранговые соотношения этих случайных величин. Абсолютные значения величин $X_1, \ldots, X_d$ используются только при оценивании их одномерных распределений.

Для решения задачи классификации таксонов необходимо знать апостериорную вероятность (4.1). Эта вероятность выражается через частную производную функции копулы C, о чем утверждает следующая теорема.

Теорема 4.3. *Пусть X, Y — две случайные величины с одномерными функциями распределения $F(X)$, $G(Y)$. Тогда условная вероятность $P(Y \leqslant y|X = x)$ равна частной производной копулы:*

$$P(Y \leqslant y|X = x) = \frac{\partial}{\partial v}C(u, v)|_{(G(y),F(x))},$$

взятой в точке

$$u = G(y), \quad v = F(X).$$

В нашей задаче имеется n случайных величин, соответствующих признакам, и случайная величина Y.

Для оценки условной вероятности необходимо ввести некоторые дополнительные обозначения.

Имеется набор объектов $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$. Каждый объект описывается n признаками. Обозначим одномерные функции распределения y и всех компонент многомерной случайной величины $\mathbf{x}$:

$$G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n).$$

Обозначим совместные функции распределения упорядоченных поднаборов - векторов $\mathbf{x}^k = (x^1, \ldots, x^k)$ размерности от 1 до n:

$$F^k_{\mathbf{X}^k}(\mathbf{x}^k), \quad \mathbf{x}^k = (x^1, \ldots, x^k), \quad k = 1, \ldots, n.$$

Для нахождения условной вероятности $P(Y \leqslant y|\mathbf{x}_i)$, воспользуемся частной производной копулы $C(u, v)$ по переменной u:

$$P(Y \leqslant y|\mathbf{x}_i) = \frac{\partial}{\partial u}C(u, v)|_{F^n_{\mathbf{X}^n}(\mathbf{x}^n_i),G_Y(y)},$$

взятой в точке

$$u = F^n_{\mathbf{X}^n}(\mathbf{x}^n_i), \quad v = G_Y(y).$$

Неизвестной в этой формуле является функция совместного распределения $F^n_{\mathbf{X}^n}$. Чтобы найти эту функцию, воспользуемся теоремой 3:

$$F^n_{\mathbf{X}^n}(\mathbf{x}^n) = C^{n-1}(u, v)|_{F^{n-1}_{\mathbf{X}^{n-1}}(\mathbf{x}^{n-1}),G^n_{X_n}(x_n)},$$

$$\ldots,$$

$$F^i_{\mathbf{X}^i}(\mathbf{x}^i) = C^{i-1}(u, v)|_{F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}),G^i_{X_i}(x_i)},$$

$$\dots,$$

$$F^2_{X^1,X^2}(x^1,x^2) = C^1(u,v)|_{G^1_{X^1}(x^1),G^2_{X^2}(x^2)}.$$

Таким образом, чтобы оценить апостериорную вероятность $P(Y \leqslant y|\mathbf{x}_i)$, необходимо оценить все $n+1$ одномерные распределения y и компонент случайного вектора $\mathbf{x}$, а также n копул $C, C^1, \dots, C^{n-1}$.

Копулы, используемые при построении интегрального индикатора. Для решения задачи (4.1) предлагается использовать Архимедовскую копулу:

Определение 4.2. *Копула $C(u_1,\dots,u_d)$ назыввется архимедовской, если для нее выполнены следующие условия:*

$$C(u_1,\dots,u_d) = \psi(\psi^{-1}(u_1)+\dots+\psi^{-1}(u_d)),$$

где функция ψ называется генератором, и для нее должны быть выполнено:

$$(-1)^k\psi^{(k)}(x) \geqslant 0$$

для всех $x \geqslant 0$ и $k = 0,1,\dots,d-2$. А также, функция

$$(-1)^{d-2}\psi^{d-2}(x)$$

должны быть невозрастающей и выпуклой.

Будем использовать частные случаи Архимедовской копулы, задаваемые следующими функциями-генераторами:

копула Клейтона,

$$\psi(t) = (1+\theta t)^{-\frac{1}{\theta}}, \quad \theta \in \Theta = (0,\infty)$$

и копула Гумбеля,

$$\psi(t) = \exp(-t^{\frac{1}{\theta}}), \theta \in \Theta = [1,\infty).$$

Отметим, что эти семейства копул зависят только от одного параметра θ, что значительно упрощает задачу в вычислительном смысле.

В случае копулы Гумбеля, частная производная имеет следующий вид:

$$\frac{\partial}{\partial u}C(u,v) = \left(\frac{\ln u}{\ln C}\right)^{\theta-1}\frac{C}{u}.$$

Оценка параметров копулы. Как было сказано выше, для оценки параметра $\theta \in \Theta$ копулы используются не сами случайные величины X, Y, а последовательности рангов этих величин. Выборкам X и Y соответствуют последовательности рангов:

$R_x = (R_{x_1}, \ldots, R_{x_m})$, где $R_{x_i} -$ ранг $i-$го объекта в вариационном ряду выборки X,

$R_y = (R_{y_1}, \ldots, R_{y_n})$, где $R_{y_i} -$ ранг $i-$го объекта в вариационном ряду выборки Y.

Отметим, что наиболее часто используемым методом оценки параметров распределения является метод максимизации правдоподобия, который в случае копул записывается следующим образом:

$$L(\theta) = \sum_{i=1}^{m} \log\Big(c_\theta\big(F(X_i), G(Y_i)\big)\Big),$$

$$c_\theta(u, v) = \frac{\partial^2}{\partial u \partial v} C_\theta(u, v).$$

Вместо значений функций одномерных распределений $F(X_i), G(Y_i)$ можно подставить их эмпирические значения, получив таким образом функцию псевдоправдоподобия [**?**]:

$$L^{'}(\theta) = \sum_{i=1}^{m} \left(\log c_\theta \left(\frac{R_i}{m+1}, \frac{S_i}{m+1} \right) \right).$$

Заметим, что функция L' зависит только от самой копулы C_θ, то есть, в нашем случае, только от параметра θ, и максимизация этой функции не представляет собой большой вычислительной сложности.

Благодаря этому способу, задача оценки распределений $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$ распадается на два независимых этапа: оценка параметра θ_i копул C^i путем максимизации псевдоправдоподобия и оценка параметров одномерных распределений $G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n)$ с помощью метода максимума правдоподобия.

Алгоритм оценки апостериорного распределения. Приведем подробный алгоритм оценки распределений $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$. Как было сказано выше, необходимо оценить $n+1$ одномерное распределение $G^0_Y(y), G^1_{X^1}(x^1), \ldots, G^n_{X^n}(x^n)$ и n функций копулы $C, C^1, \ldots, C^n$.

1. Оцениваются одномерные распределения $G^1_{X^1}(x^1)$, $G^2_{X^2}(x^2)$. Все функции $G^i_{X^i}(x^i)$ будем искать в классе бета-распределений. То есть, распределение случайной

величины X задается плотностью вероятности g_X, имеющей вид:

$$\begin{cases} g_X(x) & = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \\ B(\alpha,\beta) & = \int\limits_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx. \end{cases}$$

Параметры α и β для этого распределения оцениваются методом моментов. Для этого численно решается система уравнений:

$$\begin{cases} E(X) & = \frac{\alpha}{\alpha+\beta}, \\ D(X) & = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{cases}$$

2. Оценим копулу $C^1(u,v)$, связывающую переменные x^1 и x^2, максимизируя функцию псевдоправдоподобия:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L'_{12}(\theta) = \sum_{i=1}^{m} \left(\log c_\theta \left(\frac{R_{x_1}}{m+1}, \frac{R_{x_2}}{m+1} \right) \right).$$

3. Оценив одномерные распределения $G^1_{X^1}(x^1)$, $G^2_{X^2}(x^2)$ и копулу $C^1(u,v)$, получаем оценку функции совместного распределения $F^2_{X^1,X^2}(x^1,x^2)$. Повторяем шаги 1-2, каждый раз прибавляя по одному новому признаку x^i и оценивая на шаге 3 функцию $F^i_{\mathbf{X}^i}(\mathbf{x}^i)$.

4. повторив n раз шаги 1-2, получим функцию совместного распределения всех признаков $F^n_{\mathbf{X}^n}$. На последнем шаге оценим функцию распределения $G^0_Y(y)$, копулу $C(u,v)$, связывающую Y и X, и найдем $\hat{y}$, доставляющий максимум апостериорной вероятности:

$$\hat{y} = \arg\max_y P(Y \leqslant y|\mathbf{x}_i) = \arg\max_y \frac{\partial}{\partial u} C(u,v)|_{F^n_{\mathbf{X}^n}(\mathbf{x}^n_i), G_Y(y)},$$

взятой в точке

$$u = F^n_{\mathbf{X}^n}(\mathbf{x}^n_i), \quad v = G_Y(y),$$

где

$$F^i_{\mathbf{X}^n}(\mathbf{x}^i_i) = C^{i-1}(u,v)|_{F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}), G^i_{X_i}(x_i)},$$

взятой в точке

$$u = F^{i-1}_{\mathbf{X}^{i-1}}(\mathbf{x}^{i-1}), \quad v = G^i_{X_i}(x_i)$$

для всех

$$i = 2, \ldots, n.$$

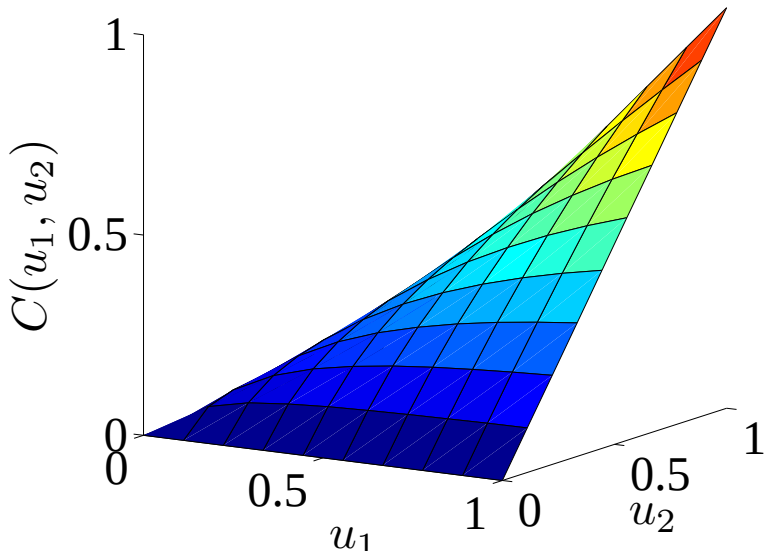


Рис. 1: Копула Клейтона

Выбор признаков. Для повышения качестве классификации на контрольной подвыборке, предлагается выбрать наиболее информативные признаки. Множество индексов признаков, включенных в функцию вероятности 4.1, назовем активным набором и обозначим $\mathcal{A} \subseteq \mathcal{J}$.

Для того, чтобы выбрать наиболее информативные признаки, предлагается использовать следующий эвристический алгоритм. Информационными будем читать те признаки, которые имеют наибольшую ранговую связь со случайной величиной Y. Чтобы понять, какие признаки имеют наибольшую связь, рассмотрим некоторые свойства копул о ранговой связи.

Утверждение 4.1. *Случайные величины X и Y являются независимыми тогда и только тогда, когда*

$$C(u, v) = uv, \quad u, v \in [0, 1],$$

где

$$C(F(x), G(y)) = H(x, y),$$

где $H(x, y)$ — совместная функция распределения случайных величин X и Y.

Утверждение 4.2. *Границы Фреше для копулы:*

$$W(u, v) \leqslant C(u, v) \leqslant M(u, v), \quad u, v \in [0, 1],$$

где

$$W(u,v) = \max(0, u+v-1)$$

— минимальная копула,

$$M(u,v) = \min(u,v)$$

— максимальная копула.

Причем, если $C(u,v) = W(u,v)$, то Y— монотонно убывающая функция X, если $C(u,v) = M(u,v)$, то Y— монотонно возрастающая функция X.

Для примера, рассмотрим копулу Гумбеля (4.1):

$$C_\theta(u,v) = \exp\left[\left((-\log(u))^\theta + (-\log(v))^\theta\right)^{\frac{1}{\theta}}\right] \quad \theta \geqslant 1.$$

При стремлении параметра копулы $\theta \to 1$, $C_\theta(u,v) \to uv$, то есть, случайные величины являются независимыми. При стремлении параметра $\theta \to \infty$, ранговая связь между случайными величинами возрастает. Таким образом, ранговая связь изменяется монотонно при варьировании параметра копулы. Для решения задачи отбора признаков будем отбирать те из них, для которых параметр копулы со случайной величиной Y является наибольшим.

Исходя из этого рассуждения, предлагается следующий алгоритм.

1. Примем пустое множество активных признаков

$$\mathcal{A} = \emptyset.$$

2. Для всех $j = 1, \ldots, n$ вычислим параметры θ_j для копул $C_{\theta_j}(F_i(x^j), G(y))$ и включим в набор

$$\mathcal{A} = \mathcal{A} \cup \{k\}$$

тот признак k, для которого

$$k = \arg\max_{j \in \mathcal{J}} \theta_j.$$

Обозначим множество оставшихся признаков

$$\mathcal{J}' = \mathcal{J} \setminus \mathcal{A}.$$

3. Для всех признаков $j \in \mathcal{A}$ и всех $k_1, \ldots, k_\mathcal{A}$ вычислим параметры θ_j для копул

$$C_{\theta_j}(F_i(x_i), H_{k_1,\ldots,k_\mathcal{A}}(x^{k_1}, \ldots, x^{k_\mathcal{A}}))$$

и включим в набор

$$\mathcal{A} = \mathcal{A} \cup \{k\}$$

тот признак k, для которого

$$k = \arg\max_{j \in \mathcal{J}'} \theta_j.$$

4. Будем повторять шаг 3, пока значение ошибки на контрольной выборке не стабилизируется.

4.2 Алгоритм криволинейной регрессии

Постановка задачи. Криволинейная модель $f(\mathbf{w}, \mathbf{x}_i)$ имеет вид

$$f(\mathbf{w}, \mathbf{x}_i) = \xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big), \tag{4.1}$$

$$h(\mathbf{w}, \mathbf{x}_i) = \sum_{j \in \mathcal{J}} u_j g(\mathbf{b}_j, x_{ij}). \tag{4.2}$$

где вектор параметров $\mathbf{w} = [\mathbf{b}_0; \mathbf{b}_1; \ldots; \mathbf{b}_n; \mathbf{u}] = [\mathbf{b}_0^T, \mathbf{b}_1^T, \ldots, \mathbf{b}_n^T, \mathbf{u}^T]^T$ состоит из векторов $\mathbf{b}_j$ — параметров монотонной коррекции j-го признака χ_j и весовых коэффициентов признаков $\mathbf{u} = [u_1, \ldots, u_j, \ldots, u_n]^T$. Функция g монотонной коррекции задана следующим образом:

$$g(j, \boldsymbol{\chi}) : \boldsymbol{\chi} \mapsto \mathbf{b}_j = \begin{cases} 1 \mapsto b_{j1}, \\ 2 \mapsto b_{j2}, \\ \ldots \\ k_j \mapsto b_{jk_j}. \end{cases}$$

При этом соблюдается условие монотонности параметров,

$$\mathrm{Ord}(\mathbf{b}_j) : 0 < b_{j1} < b_{j2} < \cdots < b_{jk_j} < 1 \quad \text{для} \quad j = 1, ..., n \quad \text{и} \tag{4.3}$$

$$\mathrm{Ord}(\mathbf{b}_0) : b_{01} < b_{02} < \cdots < b_{0k_0}.$$

Функция $\xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big)$ определяет для числа $h(\mathbf{w}, \mathbf{x}_i)$ ближайшую по модулю компоненту вектора $\mathbf{b}_0$:

$$\xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big) = \arg\min_{j \in \mathcal{J}} |b_{0j} - h(\mathbf{w}, \mathbf{x}_i)|.$$

Введя обозначение для матрицы скорректированных экспертных оценок

$$G = [g_{ij}] = [g(\mathbf{b}_j, x_{ij})], \quad i \in \mathcal{I}, j \in \mathcal{J},$$

перепишем (4.1) и (4.2) в виде модели интегрального индикатора

$$f(\mathbf{w}, \mathbf{x}_i) = \xi(\mathbf{b}_0, [G\mathbf{u}]_i). \tag{4.4}$$

Назначим функцией ошибки модели сумму квадратов регрессионных остатков,

$$S(\mathbf{w}) = \|\mathbf{f}(\hat{\mathbf{w}}, X) - \mathbf{y}\|_2^2 + \lambda\|\hat{\mathbf{u}}\|_2^2,$$

включающую регуляризующее слагаемое с фиксированным коэффициентом λ, где $\hat{\mathbf{w}}$ и $\hat{\mathbf{u}}$ — параметры, которые необходимо оценить.

Оценивание параметров модели. Оценивание параметров $\mathbf{w}$ модели $\mathbf{f}$ выполняется итеративно. Перед началом итераций значения векторов $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_n$ назначены таким образом, что функция g является тождественной, $g = \mathrm{id}$. Оценивание параметров выполняется в три этапа. Сначала при фиксированных значениях векторов $\hat{\mathbf{b}}_0, \ldots, \hat{\mathbf{b}}_n$ оцениваются весовые коэффициенты

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}\in\mathbb{R}^n} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]\big).$$

Затем при фиксированных значениях коэффициентов $\hat{\mathbf{u}}$ оцениваются параметры монотонной коррекции

$$[\mathbf{b}_1; ...; \mathbf{b}_n] = \arg\min_{\mathrm{Ord}(\mathbf{b}_1),\ldots,\mathrm{Ord}(\mathbf{b}_n)} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]\big)$$

с учетом требования монотонности значений этих параметров. На последнем этапе оценивается вектор $\mathbf{b}_0$

$$\mathbf{b}_0 = \arg\min_{\mathrm{Ord}(\mathbf{b}_0)} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]\big).$$

Итерации выполняются до стабилизации функции ошибки S.

Рассмотрим эти три этапа более подробно. За начальное приближение примем столбцы матрицы G

$$\hat{G} = [\mathbf{g}(\hat{\mathbf{b}}_1, \boldsymbol{\chi}_1), \ldots, \mathbf{g}(\hat{\mathbf{b}}_n, \boldsymbol{\chi}_n)] = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n],$$

поскольку, как было сказано выше, $g = \mathrm{id}$, и вектор $\hat{\mathbf{y}} = \mathbf{y}$. Таким образом, векторы $\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_n$ в начальном приближении в качестве элементов содержат элементы множеств $\mathbb{L}_0, \mathbb{L}_1, \ldots, \mathbb{L}_n$.

Шаг 1. Найдем $\hat{\mathbf{u}}$ при фиксированных $\hat{\mathbf{b}}_0, \ldots, \hat{\mathbf{b}}_n$:

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\hat{\mathbf{y}} - \hat{G}\mathbf{u}\| + \lambda\|\mathbf{u}\|.$$

Решение на шаге 1 имеет вид:

$$\hat{\mathbf{u}} = (\hat{G}^T\hat{G} + \lambda I)^{-1}\hat{G}^T\hat{\mathbf{y}}.$$

Шаг 2. При фиксированных $\hat{\mathbf{b}}_0, \hat{\mathbf{u}}$ оценим скорректированную матрицу описаний

$$G = [\mathbf{g}(\mathbf{b}_1, \boldsymbol{\chi}_1), \ldots, \mathbf{g}(\mathbf{b}_n, \boldsymbol{\chi}_n)] = [\mathbf{g}_1, \ldots, \mathbf{g}_n].$$

Для каждого $\mathbf{g}_j \in \mathbb{R}^m$ будем вычислять вектор $\hat{\mathbf{g}}_j$, являющийся монотонной коррекцией исходного вектора $\mathbf{g}_j$:

$$\begin{cases} [\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n] = \arg\min \|\boldsymbol{\xi}(\mathbf{b}_0, G\hat{\mathbf{u}}) - \hat{\mathbf{y}}\|_2^2, \\ \text{из} \quad g_{ij_1} \leqslant g_{ij_2} \quad \text{следует} \quad \hat{g}_{ij_1} \leqslant \hat{g}_{ij_2} \quad i \in \mathcal{I}, j_1, j_2 \in \mathcal{J}, \\ g_{ij} \in [0,1] \quad i \in \mathcal{I}, j \in \mathcal{J}, \text{ согласно (4.3)}. \end{cases}$$

По векторам $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n$ затем однозначно восстанавливаются векторы $\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_n$ как упорядоченные векторы, содержащие различные элементы $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n$. Для решения этой задачи используется алгоритм градиентного спуска, описанный в [20].

Шаг 3. Наконец, при фиксированных $_1, \ldots, \hat{\mathbf{b}}_n, \hat{\mathbf{u}}$ оценим вектор $\mathbf{b}_0$ и $\hat{\mathbf{y}} = \boldsymbol{g}(\mathbf{b}_0, \mathbf{y})$:

$$\hat{\mathbf{b}}_0 = \arg \min_{\text{Ord}(\mathbf{b}_0)} \|\boldsymbol{\xi}(\mathbf{b}_0, \hat{G}\hat{\mathbf{u}}) - \boldsymbol{g}(\mathbf{b}_0, \mathbf{y})\|_2^2.$$

Выбор признаков при классификации. Множество индексов признаков, включенных в модель, назовем активным набором и обозначим $\mathcal{A} \subseteq \mathcal{J}$.

Поставим задачу выбора наиболее информативных признаков следующим образом. Разобъем выборку $\mathcal{D}$ на две подвыборки, обучающую и тестовую. Обозначим индексы элементов этих подвыборок сооветственно $\mathcal{L} \sqcup \mathcal{T} = \mathcal{I}$. Для некоторого активного набора признаков $\mathcal{A}$ найдем на обучающей подвыборке $\mathcal{D}_\mathcal{L}$ оптимальные, согласно заданной функции ошибки S, параметры $\hat{\mathbf{w}}_\mathcal{A}$,

$$\hat{\mathbf{w}}_\mathcal{A} = \arg\min_{\mathbf{w}} S(\mathbf{w}_\mathcal{A} | \mathcal{D}_\mathcal{L}).$$

Затем выберем наиболее информативные признаки — активный набор $\hat{\mathcal{A}}$ по всем поднаборам индексов признаков $\mathcal{A} \subseteq \mathcal{J}$, доставляющий на тестовой выборке $\mathcal{D}_\mathcal{L}$ минимум функции ошибки:

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(\hat{\mathbf{w}}_\mathcal{A} | \mathcal{D}_\mathcal{T}).$$

Для выбора наиболее информативного подмножества признаков используется итеративный алгоритм добавления признаков.

На первом шаге этого алгоритма принимается активное множество информативных признаков $\hat{\mathcal{A}} = \emptyset$. На каждом следующем шаге к множеству $\hat{\mathcal{A}}$ добавляется признак с индексом $\hat{j}$, такой что

$$\hat{j} = \arg\min_{j \in \mathcal{J} \setminus \mathcal{A}} S(\hat{\mathbf{w}}_{\hat{\mathcal{A}} \cup \{j\}} | \mathcal{D}_{\mathcal{T}}).$$

Эта процедура продолжается итеративно до тех пор, пока значение функции ошибки S на контрольной выборке $\mathcal{D}_{\mathcal{T}}$ не достигнет минимума.

5 Вычислительные эксперимент

5.1 Реальные данные: категоризация редких видов Красной книги РФ

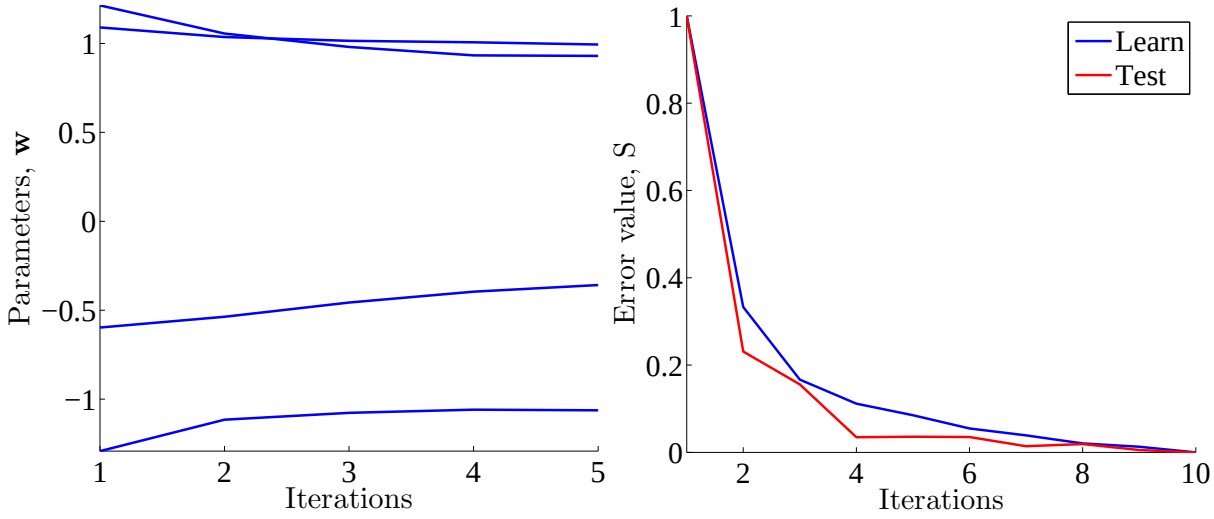
Рассматривается задача построения интегрального индикатора в ранговых шкалах. В качестве практического приложения рассматривается проблема определения статуса угрожаемых видов животных, входящих в список Красной книги РФ. В Красной книге РФ принята следующая категоризация редкости видов (таксонов) по степени угрозы их исчезновения. Имеется шесть различных категорий статуса (меток классов) таксонов: 0 — вероятно исчезнувшие, 1 — находящиеся под угрозой исчезновения, 2 — сокращающиеся в численности, 3 — редкие, 4 — неопределенные по статусу, 5 — восстанавливаемые и восстанавливающиеся. Эта категоризация является монотонной: метки классов ранжированы по возрастанию биологического разнообразия.

Каждый таксон описан набором признаков, отражающих его состояние. Эксперт, владеющий информацией о таксоне, выставляет оценку для каждого признака в ранговой шкале. Таким образом, задана матрица «объект-признак», состоящая из описаний таксонов и вектор меток классов таксонов. Требуется построить модель, восстанавливающую класс таксона из Красной книги РФ по его описанию.

Задача ревизии Красной книги РФ и построения модели вычисления интегрального индикатора является актуальной из-за постоянного пополнения книги новыми записями о таксонах.

5.2 Сходимость параметров

Алгоритм параметризованных конусов. Сходимость алгоритма в случае малого количества параметров показана на рис. 2. Видно, что значение ошибки на обучающей и контрольной выборке монотонно уменьшается. На рис. 3 показана сходимость алгоритма для случай количества признаков, сравнимого с количеством объектов. Из-за большой подгонки под обучающие данные, значение ошибки на контроле сначала уменьшается, затем начинает расти. На рис. 4 показан случай большого количества параметров. Алгоритм переобучается с первой же итерации, и ошибка на контроле начинает расти сразу.



(a) Сходимость параметров

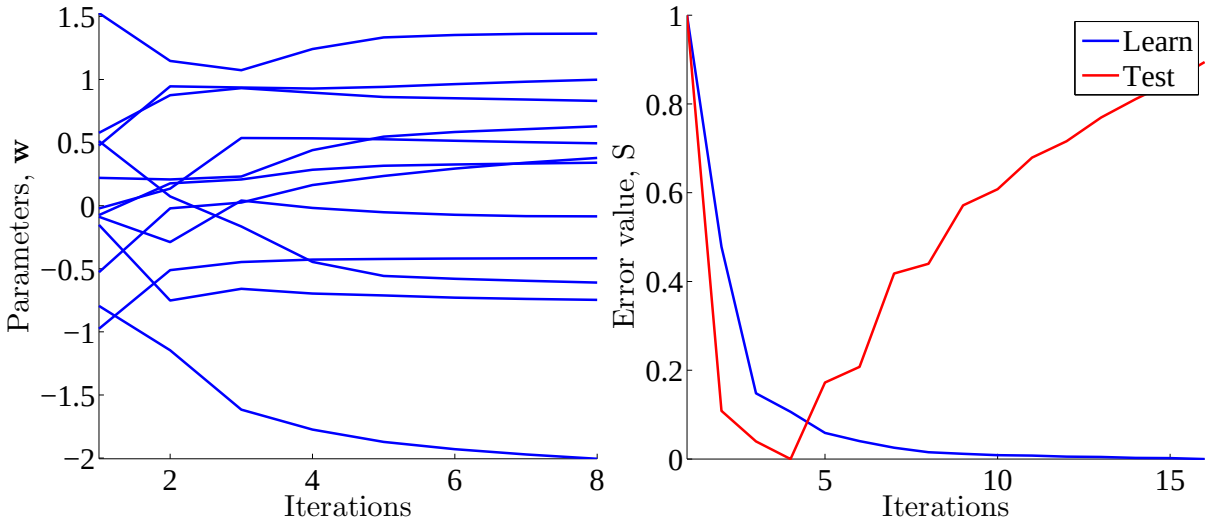
(b) Значение ошибки

Рис. 2: Случай малого количества параметров

Алгоритм криволинейной регрессии. Сходимость параметров алгоритма криволинейной регрессии показана на рис. 5, 6, 7.

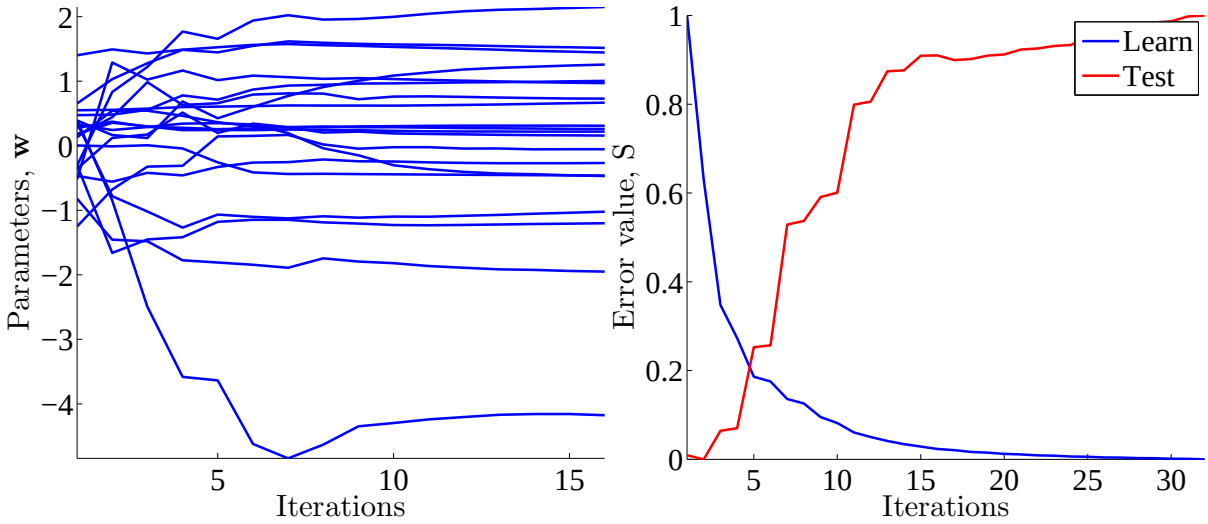
На рис. 5 показана сходимость весов регрессии $\mathbf{u}$, на рис. 6 — сходимость элементов вектора $\mathbf{b}_0$. По оси абсцисс отложено количество итераций, по оси ординат — количественное значение каждого признака. Видно, что сходимость наблюдается на десятой итерации.

На рис. 7 показана зависиость функции ошибки от количества выбираемых признаков. Видно, что ее минимум достигается при семи признаках, и значение средней ошибки равно $Q = 0.75$.



(a) Сходимость параметров (b) Значение ошибки

Рис. 3: Случай среднего количества параметров



(a) Сходимость параметров (b) Значение ошибки

Рис. 4: Случай большого количества параметров, переобучение

Статистический алгоритм. На рис. 8а показана зависимость ошибки классификации от количества выбранных признаков. Оптимальное значение достигается при $|\mathcal{A}| = 4$. В исходной таблице данных эти признаки индексированы номерами 22, 24, 23 и 20.

На рис. 8б показана зависимость параметра копулы от количества выбранных признаков. Видно, что значение параметра монотонно убывает с ростом количества признаков.

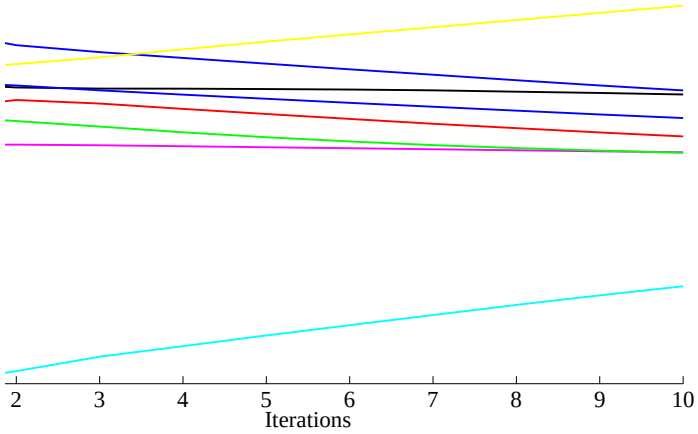


Рис. 5: Сходимость весов признаков

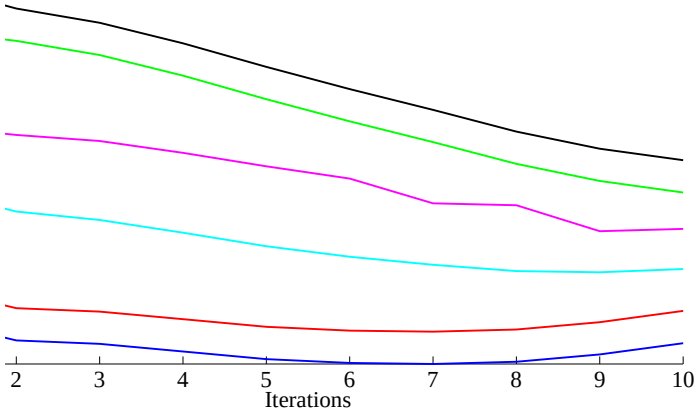


Рис. 6: Сходимость элементов вектора $\mathbf{b}_0$.

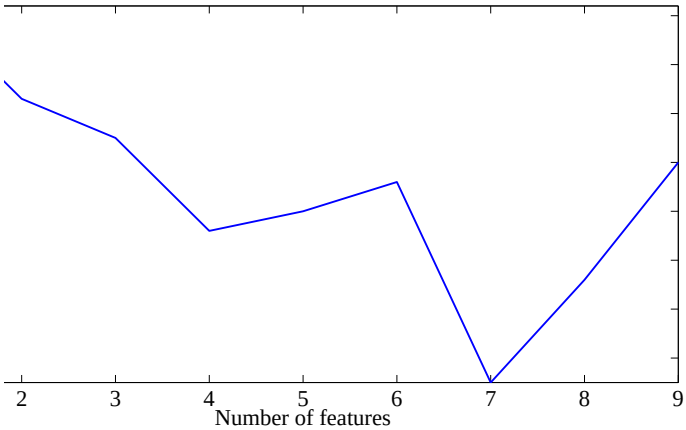
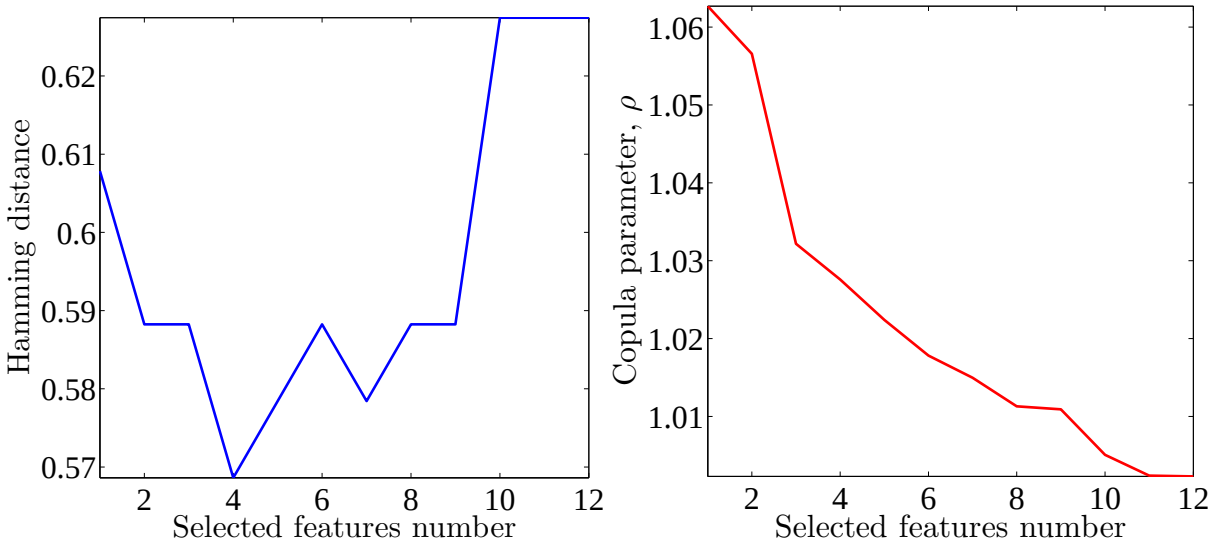


Рис. 7: Зависимость функции ошибки от количества выбираемых признаков.

Сравнение результатов. В ходе вычислительного эксперимента сравнивались результаты алгоритма, основанного на параметризации конусов, и двух альтернатив-



(a) Зависимость ошибки классификации

(b) Зависимость параметра копулы

Рис. 8: Зависимость классификации от количества выбранных признаков

ных алгоритмов.

Оценивалась ошибка на обучении и контроле, контрольная выборка выбиралась методом Leave-One-Out. Результаты эксперимента представлены в табл. 1.

Таблица 1: Сравнение алгоритмов классификации

Алгоритм	Средняя ошибка на обучении	LOO	Время построения модели, сек
Конусы	0.29	0.58	1.2
Криволинейная регрессия	0.57	0.71	3.6
Копулы	0.57	0.61	0.25

6 Заключение

В работе предложен алгоритм построения интегральных индикаторов объектов, описанных в смешанных шкалах. Предложено описание порядковой шкалы частично упорядоченным множеством. В соответствие частично упорядоченному множеству поставлен конус. Установлено соответствие между образующими этого конуса

и столбцами матрицы инцидентности графа, соответствующего частично упорядоченному множеству. Предложено параметризовать конус, задающий множество значений признака. Предложен итеративный алгоритм поиска оптимальных параметров путем минимизации функции ошибки, доказана его сходимость к оптимальному решению. Также предложен непараметрический подход, основанный на построении допустимого множества значений интегральных индикаторов как суммы конусов входящих в модель признаков.

Для оценки качества предлагаемых алгоритмов рассмотрено два альтернативных подхода к решению задачи. Первым является статистический подход, в котором взаимосвязь между порядковыми признаками устанавливается функцией копулы. Второй подход является обобщением линейной регрессии на случай порядковых признаков. Приведены иллюстрации сходимости параметров алгоритмов и сравнение результатов на примере категоризиации редких видов Красной книги РФ.

Список литературы

[1] А. И. Орлов. *Эконометрика.* Экзамен, 2002.

[2] Johannes Fuernkranz and Eyke Huellermeier. *Preference learning.* Springer, 2011.

[3] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. *Lecture Notes in Computer Science*, 4321:291–324, 2007.

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

[5] Eyke Huellermeierllermeier, Johannes Fuernkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1916, November 2008.

[6] Weiwei Cheng, Michael Rademaker, Bernard De Baets, and Eyke Huellermeier. Predicting partial orders: Ranking with abstention. *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science*, 6321:215–230, 2010.

[7] Frank Wilcoxon. Individual comparisons by ranking methods. *Breakthroughs in Statistics*, pages 196–202, 1992.

[8] Andrew Trotman. Learning to rank. *Information Retrieval*, 8:381, 2005.

[9] А. И. Орлов. *Нечисловая статистика.* МЗ-Пресс, 2004.

[10] Sergei Obiedkov Sergei Kuznetsov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14:189–216, 2002.

[11] П. С. Александров. *Введение в теорию множеств и общую топологию.* Наука, 1977.

[12] Gunther Schmidt. *Relational Mathematics (Encyclopedia of Mathematics and its Applications).* Cambridge University Press, 2010.

[13] Saroj B. Malik Sujit Kumar Mitra, P. Bhimasankaram. *Matrix Partial Orders, Shorted Operators and Applications.* 2010.

[14] М.П. Кузнецов and В.В. Стрижов. Построение интегрального индикатора с использованием ранговой матрицы описаний. *Интеллектуализация обработки информации. Доклады 9-й международной конференции*, pages 130–132, 2012.

[15] М.П. Кузнецов. Построение интегрального индикатора в ранговых шкалах с использованием копул для анализа совместного распределения критериев. *Машинное обучение и анализ данных*, 1:411–419, 2012.

[16] М. В. Уханов. Алгоритм построения суммы многогранников. *Вестник ЮУрГУ*, pages 39–44, 2001.

[17] А. Схрейвер. Теория линейного и целочисленного программирования. *Ж. вычисл. матем. и матем. физ.*, 1:360, 1991.

[18] Н. Б. Черникова. Алгоритм для нахождения общей формулы неотрицательных решений системы линейных уравнений. *Ж. вычисл. матем. и матем. физ.*, 4:733–738, 1964.

[19] М. П. Кузнецов, В. В. Стрижов, and М.М Медведникова. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах. *Научно-технический вестник СПб ГПУ. Информатика. Телекоммуникации. Управление*, 5, 2012.

[20] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.