Построение интегрального индикатора в ранговых шкалах с использованием копул для анализа совместного распределения критериев*

Кузнецов М. П.
mikhail.kuznecov@phystech.edu
Московский физико-технический институт

Предложен метод построения интегрального индикатора на основе критериев, выставленных в ранговых шкалах. Для анализа совместного распределения критериев предложено использовать копулы. Предложен алгоритм выбора признаков, основанный на выборе копулы с наибольшим параметром. Работа проиллюстрирована задачей определения статуса редких видов, включенных в Красную книгу РФ.

Ключевые слова: *порядковые шкалы, копула, теорема Скляра, интегральные индикаторы, экспертные оценки.*

Integral indicator construction using copulas*

Kuznetsov M. P.
Moscow Institute of Physics and Technology

We construct an integral indicator of the IUCN Red List of Threatened species. Method of an integral indicator construction based on copulas which describe statistical bounds between the features. We propose a two-step algorithm of the parameters estimation. On the first step we estimate parameters of a marginal distribution of the features. On the second step we estimate copula parameters.

Keywords: *ordinal scales, copula, Sklar theorem, integral indicators, expert estimations.*

Введение

Рассматривается задача построения интегрального индикатора в ранговых шкалах. Интегральный индикатор — это число, поставленное в соответствие объекту, и рассматриваемое как оценка его качества. Интегральными индикаторами называется вектор оценок, поставленный в соответствие набору объектов.

Ранее в работах [1, 2] были описаны процедуры построения интегральных индикаторов с использованием описаний объектов в линейных шкалах. При этом интегральный индикатор являлся уточнением оценки, заданной экспертом в линейной или ранговой шкале. В данной работе рассматривается ранговое описание объектов.

Описаниями объектов являются критерии, выбранные экспертами. Для построения интегрального индикатора оценивается совместная вероятность распределения критериев. Для решения этой задачи используются копулы [3, 4] — функции, являющиеся многомерными параметрическими функциями распределения равномерно распределенных случайных величин. Для оценки параметров распределения предлагается итеративный алгоритм: на первом шаге оцениваются параметры одномерных распределений критериев, на втором шаге оцениваются параметры копул.

Научный руководитель В.В. Стрижов
Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00709.

Использование копул для оценки распределений помогает справиться с проблемой ранговости критериев. Для оценки параметра копулы необходимо знать только ранговые соотношения между величинами, а не их абсолютные значения.

Предлагается алгоритм выбора наиболее информативных критериев, использующий параметры копулы в качестве показателя информативности. Эти параметры обладают свойством монотонности: чем больше параметр копулы, тем больше ранговая связь между двумя случайными величинами. Предлагается выбирать те критерии, которые имеют наибольшую ранговую связь с экспертными оценками интегральных индикаторов.

В качестве прикладной задачи рассматривается задача определения статуса угрожаемых видов животных, входящих в список Красной книги РФ [5, 6]. В Красной книге РФ принята следующая категоризация редкости видов (таксонов) по степени угрозы их исчезновения. Имеется шесть различных категорий статуса (интегральных индикаторов) 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: интегральные индикаторы 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 интегральный индикатор 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.

Постановка задачи

Пусть X — множество объектов, Y — конечное множество меток классов. Множество $X \times Y$ является вероятностным пространством с совместной функцией распределения $P(x, y)$.

Задано множество $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}, \quad i \in \mathcal{I} = \{1, \ldots, m\}$, которое является выборкой пар $(\mathbf{x}_i, y_i)$. Объект $\mathbf{x}_i \in X$— таксон, $y_i \in Y$ — метка класса, соответствующая этому таксону.

Описание объекта $x_i = [\boldsymbol{\chi}^1, \ldots, \boldsymbol{\chi}^j, \ldots, \boldsymbol{\chi}^n]^\intercal, \quad j \in \mathcal{J} = \{1, \ldots, n\}$ — это набор экспертных оценок признаков. Оценки объектов по признакам выставлены в ранговых шкалах. Каждый признак χ_j имеет собственную ранговую шкалу $\mathbb{L}_j$, состоящую из k_j упорядоченных элементов $\mathbb{L}_j = \{1 \prec 2 \prec \cdots \prec k_j\}$. Значение класса y также принадлежит упорядоченному множеству $\mathbb{L}_0 = \{1 \prec 2 \prec \cdots \prec k_0\}$.

Решается задача классификации объектов. Для этого 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 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 $[u_1, \ldots, u_d]^\intercal$, 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 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 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 (2). Эта вероятность выражается через частную производную функции копулы C, о чем утверждает следующая теорема.

Теорема 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:

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

$$\ldots,$$

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

$$\ldots,$$

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

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 $C, C^1, \ldots, C^{n-1}$.

Копулы, используемые при построении интегрального индикатора

Для решения задачи (2) предлагается использовать Архимедовскую копулу:

Определение 2. *Копула $C(u_1, \ldots, u_d)$ называется архимедовской, если для нее выполнены следующие условия:*

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

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

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

для всех $x \geqslant 0$ и $k = 0, 1, \ldots, d-2$. 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.

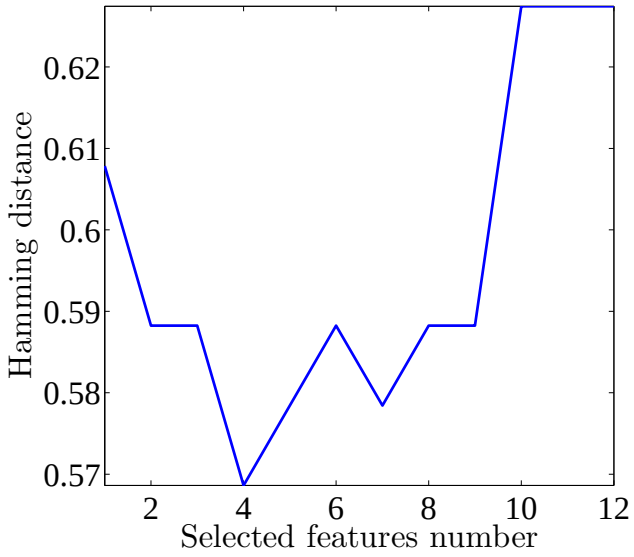

Рис. 1: Зависимость ошибки классификации от количества выбранных признаков

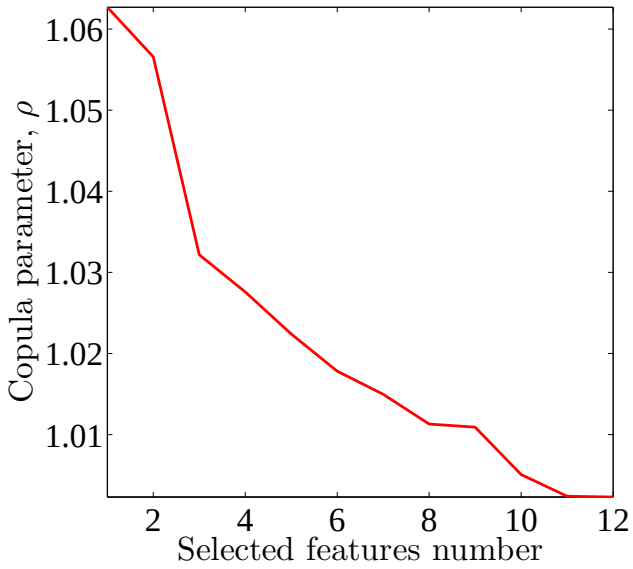

Рис. 2: Зависимость параметра копулы от количества выбранных признаков

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

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

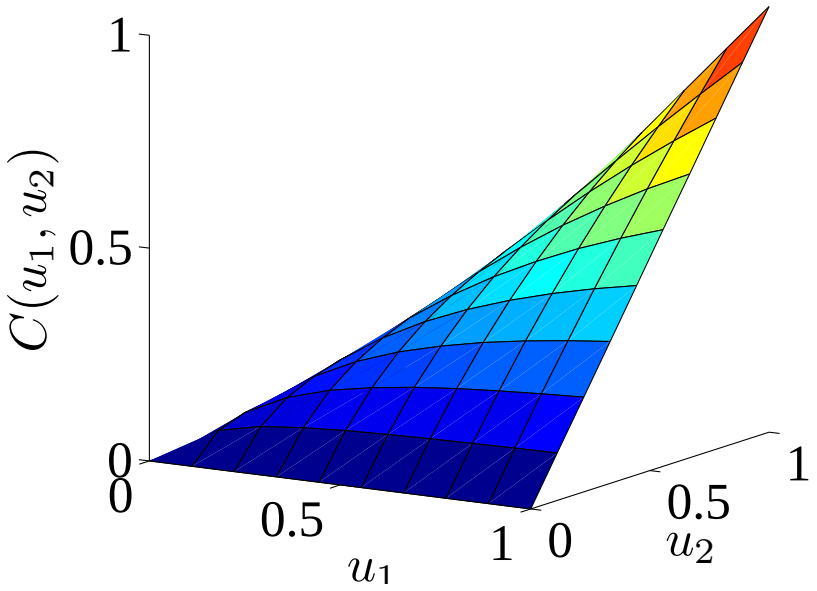

Рис. 3: Копула Клейтона

Выбор признаков

Так как число объектов в данной задаче, определенное составом Красной книги РФ, сопоставимо с числом признаков, необходимо выбрать наиболее информативные признаки. Множество индексов признаков, включенных в функцию вероятности 2, 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 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 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 (2):

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

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

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

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.

Вычислительный эксперимент

Работа алгоритма иллюстрируется данными из Красной Книги РФ. Экспертами заполнена таблица данных для 29 различных объектов. Каждый объект описывается 102 признаками.

На рис. 1 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. 2 показана зависимость параметра копулы от количества выбранных признаков. Видно, что значение параметра монотонно убывает с ростом количества признаков.

Литература

[1] Стрижов В. В. Уточнение экспертных оценок с помощью измеряемых данных // Заводская лаборатория. Диагностика материалов. 2006, Т. 72(7). С. 59–64.

[2] Strijov V., Granic G., Juric J., Jelavic B., Maricic S.A. Integral indicator of ecological impact of the Croatian thermal power plants // Energy, 2011. Vol. 36(7). Pp. 4144–4149.

[3] Roger B. Nelsen An Introduction to Copulas // Springer, 1998

[4] Edward W. Frees and Emiliano A. Valdez Understanding relationships using copulas // North american actuarial journal, 2012. Vol. 2. Pp. 104-141.

[5] 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.

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

[7] 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.

[8] Литвак Б. Г. Экспертная информация: Методы получения и анализа. М.: Радио и связь, 1982. С. 69–88.

[9] Орлов А. И. Организационно-экономическое моделирование: часть 2. Экспертные оценки. М: МГТУ им. Н. Э. Баумана, 2011. 486 с.

[10] Boyd S. and Vandenberghe L. Convex Optimization // Cambridge University Press. 2004.