Kuznetsov M.P., Strijov V.V., Medvednikova M.M. Multiclass classification of objects with the rank-scale description // Notices on Science and Technology of SPb. PSU, 2012, 5 : 92-95.

УДК 519.584

М.П. Кузнецов, В.В. Стрижов, М.М. Медведникова

АЛГОРИТМ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ ОБЪЕКТОВ, ОПИСАННЫХ В РАНГОВЫХ ШКАЛАХ

Рассматривается задача построения интегрального индикатора в ранговых шкалах [1,2,3]. В качестве практического приложения рассматривается проблема определения статуса угрожаемых видов животных, входящих в список Красной книги РФ [4]. В Красной книге РФ принята следующая категоризация редкости видов (таксонов) по степени угрозы их исчезновения. Имеется шесть различных категорий статуса (меток классов) таксонов: вероятно исчезнувшие, находящиеся под угрозой исчезновения, сокращающиеся в численности, редкие, неопределенные по статусу, восстанавливающиеся. Эта категоризация является монотонной: метки классов ранжированы по возрастанию биологического разнообразия. Назначение категории таксона может быть выполнено одним из методов согласования экспертных оценок [7,8], или путем аналитического вычисления категории на основе его описания, с учетом предложенной аналитиками модели [9].

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. Ранее были предложен ряд алгоритмов решения данной задачи [3,7]. Так как эти решения не включают процедуру выбора наиболее информативных признаков, ниже предлагается

поставить и решить задачу нелинейной коррекции экспертных оценок.

Постановка задачи

Задана множество $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}, \quad i \in \mathcal{I} = \{1, \dots, m\}$ пар. Каждая пара состоит из описания объекта $\boldsymbol{x}_i$ (таксона) и соответствующей ему метки класса y_i (категория статуса таксона).

Описание объекта $\boldsymbol{x} = [\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_j, \dots, \boldsymbol{\chi}_n], \quad j \in \mathcal{J} = \{1, \dots, n\}$ — это набор экспертных оценок признаков. Оценки объектов по признакам выставлены в ранговых шкалах. Каждый признак i_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 $\mathbb{L}_0 = \{1 \prec 2 \prec \cdots \prec k_0\}$.

Рассмотрим постановку задачи многоклассовой классификации в ранговых шкалах, включающую криволинейную модель $f(\boldsymbol{w}, \boldsymbol{x}_i)$ и соответствующую ей вектор-функцию $\boldsymbol{f}(\boldsymbol{w}, X) = [f(\boldsymbol{w}, \boldsymbol{x}_1), \dots, f(\boldsymbol{w}, \boldsymbol{x}_m)]$ с матрицей описаний $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_m T]$ и зависисмой переменной $\boldsymbol{y} = [y_1, \dots, y_i, \dots, y_m]$, где $\boldsymbol{w} = [w_1, \dots, w_s]$ — параметры модели. Эта модель должна доставлять минимум заданной функции ошибки $S(\boldsymbol{f}(\boldsymbol{w}, X), \boldsymbol{y})$.

Криволинейная модель $f(\boldsymbol{w}, \boldsymbol{x}_i)$ имеет вид

$$f(\boldsymbol{w}, \boldsymbol{x}_i) = \xi(\boldsymbol{b}_0, h(\boldsymbol{w}, \boldsymbol{x}_i)), \tag{1}$$

$$h(\boldsymbol{w}, \boldsymbol{x}_i) = \sum_{j \in \mathcal{J}} u_j \, g(\boldsymbol{b}_j, x_{ij}). \tag{2}$$

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

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

При этом соблюдается условие монотонности параметров,

$$\mathrm{Ord}(\boldsymbol{b}_j)\colon 0 < b_{j1} < b_{j2} < \cdots < b_{jk_j} < 1 \quad \text{для} \quad j = 1,\ldots,n \quad \text{и} \quad \mathrm{Ord}(\boldsymbol{b}_0)\colon b_{01} < b_{02} < \cdots < b_{0k_0}. \tag{3}$$

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

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

Введя обозначение для матрицы скорректированных экспертных оценок

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

перепишем (1) и (2) в виде модели интегрального индикатора

$$f(\boldsymbol{w}, \boldsymbol{x}_i) = \xi(\boldsymbol{b}_0, [G\boldsymbol{u}]_i). \tag{4}$$

Назначим функцией ошибки модели сумму квадратов регрессионных остатков, $S(\boldsymbol{w}) = \| \boldsymbol{f}(\widehat{\boldsymbol{w}}, X) - \boldsymbol{y} \|_2^2 + \lambda \| \widehat{\boldsymbol{u}} \|_2^2$, включающую регуляризующее слагаемое с фиксированным коэффициентом λ, где $\widehat{\boldsymbol{w}}$ и $\widehat{\boldsymbol{u}}$ — параметры, которые необходимо оценить.

Оценивание параметров модели

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

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

Затем при фиксированных значениях коэффициентов $\widehat{\boldsymbol{u}}$ оцениваются параметры монотонной коррекции

$$[\boldsymbol{b}_1; \ldots; \boldsymbol{b}_n] = \arg\min_{\mathrm{Ord}(\boldsymbol{b}_1),\ldots,\mathrm{Ord}(\boldsymbol{b}_n)} S([\widehat{\boldsymbol{b}}_0; \ldots; \widehat{\boldsymbol{b}}_n; \boldsymbol{u}])$$

с учетом требования монотонности (3) значений этих параметров. На последнем этапе оценивается вектор $\boldsymbol{b}_0$: $\boldsymbol{b}_0 = \arg\min_{\mathrm{Ord}(\boldsymbol{b}_0)} S([\widehat{\boldsymbol{b}}_0; \ldots; \widehat{\boldsymbol{b}}_n; \boldsymbol{u}])$.

Итерации выполняются до стабилизации функции ошибки S.

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

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

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

Шаг 1. Найдем $\hat{\boldsymbol{u}}$ при фиксированных значениях $\hat{\boldsymbol{b}}_0, \dots, \hat{\boldsymbol{b}}_n$:

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

Решение на шаге 1 имеет вид: $\hat{\boldsymbol{u}} = (\hat{G}^T\hat{G} + \lambda I)^{-1}\hat{G}^T\hat{\boldsymbol{y}}$.

Шаг 2. При фиксированных $\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{u}}$ оценим скорректированную матрицу описаний $\boldsymbol{G} = [\boldsymbol{g}(\boldsymbol{b}_1, \boldsymbol{\chi}_1), \dots, \boldsymbol{g}(\boldsymbol{b}_n, \boldsymbol{\chi}_n)] = [\boldsymbol{g}_1, \dots, \boldsymbol{g}_n]$.

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

$$\begin{cases} [\hat{\boldsymbol{g}}_1, \dots, \hat{\boldsymbol{g}}_n] = \arg \min \| \boldsymbol{\xi}(\boldsymbol{b}_0, G\hat{\boldsymbol{u}}) - \hat{\boldsymbol{y}} \|_2^2, \\ \text{из} \quad g_{ij_1} \le g_{ij_2} \quad \text{следует} \quad \hat{g}_{ij_1} \le \hat{g}_{ij_2} \quad i \in \mathcal{I}, j_1, j_2 \in \mathcal{J}, \\ g_{ij} \in [0,1] \quad i \in \mathcal{I}, j \in \mathcal{J}, \text{ согласно (3).} \end{cases}$$

По векторам $\hat{\boldsymbol{g}}_1, \dots, \hat{\boldsymbol{g}}_n$ затем однозначно восстанавливаются векторы $\hat{\boldsymbol{b}}_1, \dots, \hat{\boldsymbol{b}}_n$ как упорядоченные векторы, содержащие различные элементы $\hat{\boldsymbol{g}}_1, \dots, \hat{\boldsymbol{g}}_n$. Для решения этой задачи используется алгоритм градиентного спуска.

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

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

Выбор признаков при классификации

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 модель, назовем активным набором и обозначим $\mathcal{A} \subseteq \mathcal{J}$.

Поставим задачу выбора наиболее информативных признаков следующим

образом. Разобьем выборку $\mathcal{D}$ на две подвыборки, обучающую и тестовую. Обозначим индексы элементов этих подвыборок сооветственно $\mathcal{L} \sqcup \mathcal{T} = \mathcal{J}$. Для некоторого активного набора признаков $\mathcal{A}$ найдем на обучающей подвыборке $\mathcal{D}_\mathcal{L}$ оптимальные, согласно заданной функции ошибки S, параметры $\widehat{\boldsymbol{w}}_\mathcal{A}$, $\widehat{\boldsymbol{w}}_\mathcal{A} =_{\boldsymbol{w}} S(\boldsymbol{w}_\mathcal{A}|\mathcal{D}_\mathcal{L})$.

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: $\hat{\mathcal{A}} =_{\mathcal{A}\subseteq\mathcal{J}} S(\widehat{\boldsymbol{w}}_\mathcal{A}|\mathcal{D}_\mathcal{T})$.

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

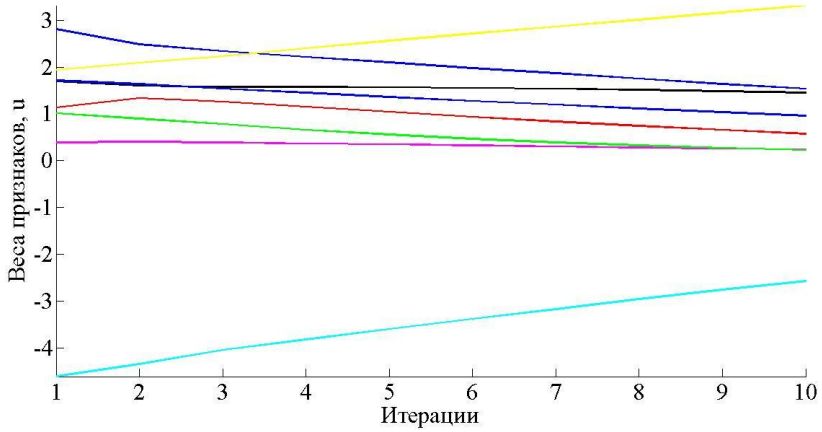

Рис. 1. Изменение весов признаков

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 $\hat{\jmath}$, такой что $\hat{\jmath} =_{j\in\mathcal{J}\setminus\mathcal{A}} S(\widehat{\boldsymbol{w}}_{\hat{\mathcal{A}}\cup\{j\}}|\mathcal{D}_\mathcal{T})$.

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

Вычислительный эксперимент

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

информативных признаков. В качестве функции ошибки классификации принимается величина

$$Q = \frac{1}{m}\sum_{i=1}^{m} S\left(f\left(\boldsymbol{w}_{\mathcal{D}\setminus i}, X_{\mathcal{I}\setminus i}\right), y_i\right),$$

где $\boldsymbol{w}_{\mathcal{I}\setminus i}$ — вектор параметров, оцененный по всей выборке без учета i-го объекта, а $X_{\mathcal{I}\setminus i}$ — матрица X, в которой исключена i-я строка. Эта метрика представляет собой среднюю ошибку классификации навсех объектах выборки, ее значение составило $Q = 0{,}75$.

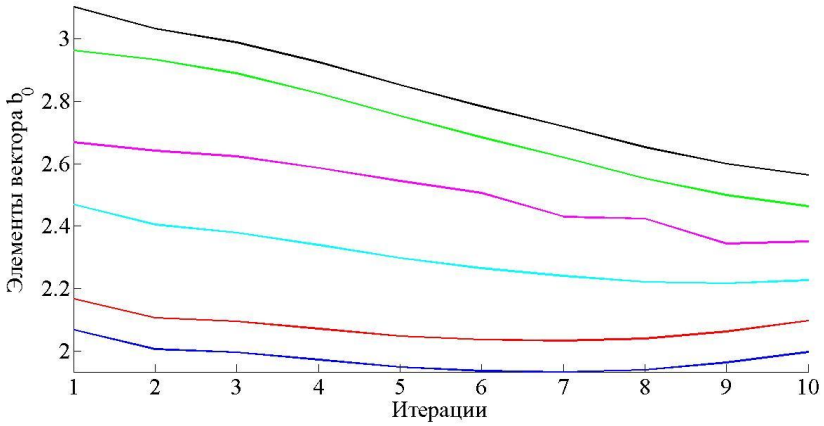

Рис. 2. Изменение элементов вектора параметров $\boldsymbol{b_0}$.

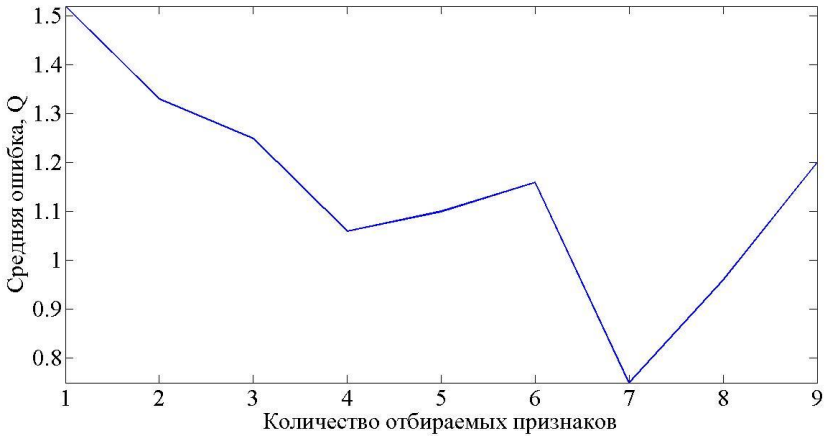

Рис. 3. Зависимость функции ошибки от количества выбираемых признаков.

Ниже представлены графики изменения параметров относительно итераций алгоритма. На рис. 1 показана изменение весов регрессии $\boldsymbol{u}$, на рис. 2 —

изменение элементов вектора $\boldsymbol{b}_0$. По оси абсцисс отложено количество итераций, по оси ординат — количественное значение каждого признака. Прекращение изменений наблюдается на десятой итерации.

На рис. 3 показана зависимость функции ошибки от количества выбираемых признаков. Видно, что ее минимум достигается при семи признаках, и значение средней ошибки равно $Q = 0.75$.

Заключение

Предложен метод построения рангового интегрального индикатора на примере задачи категоризации таксонов Красной книги РФ. Данный метод отличается от обычной задачи восстановления регрессии тем, что исходные данные представлены в ранговых шкалах и корректируются в процессе вычисления интегрального индикатора. Предложен алгоритм отбора наиболее информативных признаков. С помощью этого алгоритма получена адекватная модель получения категорий новых таксонов.

Список литературы

[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.V. et al.** Integral indicator of ecological impact of the Croatian thermal power plants // Energy. –2011. – Vol. 36(7) . – Pp. 4144–4149.

[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.

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

[5] **Литвак, Б.Г.** Экспертная информация: Методы получения и анализа. М.: Радио и связь. –1982.

[6] **Орлов, А.И.** Организационно-экономическое моделирование. Экспертные оценки. М: МГТУ им. Н. Э. Баумана. – 2011.

[7] **Kotlowski, W.** Rule learning with monotonicity constraints // Proceedings of the 26th Annual International Conference on Machine Learning. – 2009. – Vol 382. – P. 68.