На правах рукописи

Кузнецов Михаил Павлович

ПОСТРОЕНИЕ МОДЕЛЕЙ ОБУЧЕНИЯ ПО ПРЕДПОЧТЕНИЯМ С ИСПОЛЬЗОВАНИЕМ ПОРЯДКОВЫХ ЭКСПЕРТНЫХ ОЦЕНОК

05.13.18 — Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2016

Работа выполнена на кафедре интеллектуальных систем Московского физико-технического института (государственного университета)

Научный руководитель:
Стрижов Вадим Викторович, доктор физико-математических наук, Вычислительный центр им. А.А. Дородницына Российской академии наук Федерального исследовательского центра «Информатика и управление» Российской академии наук, отдел интеллектуальных систем, ведущий научный сотрудник.

Официальные оппоненты:
Дьяконов Александр Геннадьевич, доктор физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», кафедра математических методов прогнозирования факультета вычислительной математики и кибернетики, профессор;
Середин Олег Сергеевич, кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет», кафедра информационной безопасности, доцент.

Ведущая организация:
Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук.

Защита диссертации состоится “___”_________________ 2016 года в ___ часов на заседании диссертационного совета Д212.156.05 на базе Московского физико-технического института (государственного университета) по адресу: 141700, Московская обл., г. Долгопрудный, Институтский пер., д. 9, ауд. 903 КПМ.

С диссертацией можно ознакомиться в библиотеке МФТИ и на сайте института `http://mipt.ru`.

Автореферат разослан “___”_________________ 2016 года.

Ученый секретарь
диссертационного совета Федько Ольга Сергеевна

Общая характеристика работы

Диссертационная работа посвящена построению математических моделей обучения по предпочтениям. Разработанные методы опираются на анализ частично упорядоченных множеств экспертных оценок объектов и признаков.

Актуальность темы. Область обучения по предпочтениям (Fuernkranz: 2011) объединяет задачи машинного обучения с теорией группового выбора (Kendall: 1938, Kemeny: 1959) и агрегирования предпочтений для восстановления порядковой зависимости на множестве объектов с использованием линейных и порядковых экспертных оценок. В последние годы область обучения по предпочтениям приобрела особую актуальность в связи с возрастающей необходимостью решения задач информационного поиска (Liu: 2009), ранжирования, порядковой классификации и регрессии (McCullagh: 1980, Herbrich: 1999), построения монотонных композиций алгоритмов (Воронцов: 1999, Hang: 2011).

Центральной задачей работы является задача построения модели предпочтений на множестве объектов, заданных линейным и порядковым описанием. В работе рассматриваются несколько типов задач, различающихся природой описания признаков и целевой переменной, заданной на множестве объектов.

Первым типом задач является построение модели предпочтений на множестве объектов, заданных линейным признаковым описанием. Впервые задача в подобной постановке была рассмотрена для решения задачи информационного поиска (Liu: 2009). Требуется построить отображение множества объектов во множество действительных чисел, наиболее точно приближающее целевое отношение предпочтения. В задаче информационного поиска целевое отношение задается оценками асессоров на наборе документов, соответствующих запросу. Рассматривается класс линейных моделей и функция ошибки особого вида, штрафующая пары объектов, не согласованные по экспертно заданному отношению предпочтения. Моделями подобного типа являются модификация метода опорных векторов (Herbrich: 1999, Chen: 2009) и модификация бустинга (Freund: 2003). В случае, когда целевое отношение предпочтения задается конечным набором меток, решается задача порядковой классификации и регрессии (Kotlowski: 2013).

Вторым типов задач является построение модели предпочтений на множестве объектов, заданных порядковым описанием. Другими словами, на множестве объектов задан набор отношений предпочтения, соответствующих порядковым признакам. Требуется корректно определить отображение, приближающее целевое отношение предпочтения. Впервые задача в подобной постановке была рассмотрена в работе В. Коэна и Р. Шапиро (Cohen: 1999). Для восстановления линейного порядка был предложен метод оценки матрицы попарного доминирования объектов. Основным требованием к отображению является монотонность по набору порядковых признаков (Duivesteijn: 2008, Cheng: 2010). Для решения этой задачи разработаны методы построения монотонных композиций алгоритмов, рассмотренные, например, в работах К. В. Воронцова (Воронцов: 1999), а также в (Corrente: 2013). Для соблюдения условий монотонности В. В. Стрижовым был предложен подход на основе конусного представления предпочтений (Стрижов: 2011), развиваемый в данной работе.

Третьим типом задач является моделирование и согласование линейных и порядковых экспертных оценок. В данной работе рассматриваются экспертные оценки объектов, заданных линейным признаковым описанием. Методы, разрабатываемые в рамках диссертационной работы, развивают подход В. В. Стрижова (Стрижов: 2002) и используют элементы теории экспертного оценивания, исследуемой Б. Г. Литваком (Литвак: 1982) и А. И. Орловым (Орлов: 2002). В качестве частного случая задачи экспертного оценивания рассматривается задача учета экспертной информации о попарном доминирования признаков. Для ее решения используются элементы теории важности критериев, разрабатываемой В. В. Подиновским (Подиновский: 2007).

Цели работы

1. Разработка новых подходов к моделированию объектов, заданных порядковым описанием.

2. Разработка и обоснование эффективных вычислительных методов распознавания объектов, заданных линейным и порядковым описанием.

3. Реализация алгоритмов распознавания и проведение вычислительных экспериментов для проверки адекватности разработанных моделей.

Задачи работы

1. Разработать подход к моделированию порядковых признаков с использованием полиэдрального представления порядковых данных.

2. Разработать алгоритм построения модели предпочтений на множестве объектов, заданных порядковым признаковым описанием.

3. Разработать метод решения задачи порядковой классификации объектов, заданных линейным и порядковым описанием.

4. Разработать методы моделирования и согласования линейных и порядковых экспертных оценок.

5. Реализовать алгоритмы распознавания порядковых объектов и провести вычислительные эксперименты для установления границ применимости методов.

Основные положения, выносимые на защиту

1. Метод моделирования объектов, заданных порядковым описанием, с использованием полиэдрального представления предпочтений на множестве объектов.

2. Метод построения модели предпочтений объектов на основе суммы конусов предпочтений.

3. Численный метод решения задачи порядковой классификации объектов на основе оценивания матрицы попарного доминирования объектов.

4. Методы моделирования и согласования порядковых экспертных оценок объектов, заданных линейным признаковым описанием.

5. Программный комплекс, включающий в себя методы распознавания объектов, заданных линейным и порядковым описанием. Проведены вычислительные эксперименты, подтверждающие адекватность методов.

Методы исследования. Для достижения поставленных целей используется теория обучения по предпочтениям. Для исследования суперпозиции порядковых признаков используются механизмы теории голосования и общественного выбора, исследуемые в середине XX века К. Эрроу и Дж. Кемени, а также в более современных работах В. Данилова и Д. Саари. Для учета целевой переменной при построении модели предпочтений используются наработки в области обучения ранжированию, описанные в работах В. Коэна и Р. Шапиро. Исследуется обобщение задач порядковой регрессии и классификации, а также методы построения монотонных композиций, исследованные, в частности, в работах К. В. Воронцова. Для введения алгебраических структур, описывающих порядковый признак и частично упорядоченное множество, используется подход на основе конусов, впервые разработанный В. В. Стрижовым, а также теория анализа экспертной информации, исследуемая Б. Г. Литваком и А. И. Орловым. Кроме того, используются элементы теории выпуклой оптимизации, теории вероятностей, теории графов и теории важности критериев, разрабатываемой В. В. Подиновским.

Научная новизна. Разработан новый подход к построению математической модели предпочтений на основе полиэдрального представления порядковых данных. Разработаны новые вычислительные методы агрегирования порядковых признаков, основывающиеся на построении суперпозиции конусов предпочтений. Показано, что предлагаемая модель является обобщением модели порядковой классификации с монотонными ограничениями. Установлено теоретическое соответствие между конусом и матрицей инцидентности графа предпочтений. Проведено обобщение известных порядковых метрик на случай частичных порядков с использованием матрицы предпочтений.

Теоретическая значимость. Диссертационная работа устанавливает теоретические обобщения ранее полученных результатов в области обучения по предпочтениям путем введения понятия конуса предпочтений. Введенные понятия помогают провести обобщения для известных ранее понятий ранговой корреляции и монотонных ограничений, а также провести теоретический анализ свойств алгоритмов восстановления предпочтений и порядковой классификации.

Практическая значимость. Разработанные вычислительные методы показывают существенное улучшение качества решения реальных прикладных задач ранжирования и порядковой классификации. Реализованный программный комплекс позволяет проводить самостоятельные исследования в области обучения по предпочтениям и сравнивать результаты с известными методами. В рамках диссертационной работы решена практическая задача категоризации таксонов Красной книги по экспертным оценкам, предоставленным Министерством природных ресурсов РФ.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах ранжирования и порядковой классификации, публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях:

- Международная конференция «International Conference on Operational Research», 2011. *Integral Indicators and Expert Estimations of Ecological Impact.*

- Международная конференция «25th European Conference on Operational Research», 2012. *Rank-scaled Integral Indicators of Ecological Impact.*

- Международная конференция «26th European Conference on Operational Research», 2013. *The IUCN Red List threatened species categorization algorithm.*

- Международная конференция «20th Conference of the International Federation of Operational Research Societies», 2014. *Partial orders combining for object ranking problem.*

- Всероссийская конференция «Математические методы распознавания образов» ММРО-17, 2015. *Комбинирование отношений порядка для восстановления предпочтения на наборе объектов.*

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в тринадцати работах, из которых шесть опубликованы в изданиях из списка ВАК (из них три индексируемы Scopus). Список публикаций приведен в конце автореферата.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из введения, четырех разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 89 наименований. Текст работы занимает 103 страницы.

Основное содержание работы

Во введении обоснована актуальность диссертационной работы, поставлены цели и задачи исследования, аргументирована научная новизна, показана практическая значимость результатов, представлены выносимые выносимые на защиту научные положения.

В **главе 1** введены основные определения и обозначения. Сформулированы основные постановки задач в области обучения по предпочтениям, приведен обзор современных методов восстановления предпочтений. Приведено определение частично упорядоченного множества, графа и конуса предпочтений, рассмотрены их основные свойства.

Задано *множество объектов* $X = \{x_1, ..., x_m\}$ размера m. В рамках работы предполагается, что рассматриваемые объекты имеют линейное или порядковое признаковое описание.

Определение 1. *Линейное признаковое описание: объект x_i является элементом евклидова пространства $\mathbb{R}^n$.*

Определение 2. Порядковое признаковое описание*: на множестве объектов X задано n отношений предпочтения $z_1, ..., z_n$, каждое из которых описывает отношение доминирования между парами объектов, $z_j(x_i, x_k) = [x_i \succeq_j x_k]$, где $[\cdot]$ — индикаторная функция.*

Определение 3. *Целевое отношение предпочтения z_0 — отношение, описывающее искомое отношение доминирования между парами объектов из X, $z_0(x_i, x_k) = [x_i \succeq_0 x_k]$.*

Определение 4. *Модель f — отображение множества объектов X во множество действительных чисел $\mathbb{R}$.*

Для определения оптимальных параметров модели определена *функция потерь* $S(f, X, z_0)$ модели f на множестве объектов X. Основной задачей является поиск отображения $\hat{f}$, минимизирующего функцию потерь на X,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} S(f, X, z_0),$$

где $\mathcal{F}$ — класс допустимых отображений.

Определение 5. *Частичным порядком на множестве X называется бинарное отношение $\succeq$, удовлетворяющее условиям рефлексивности, антисимметричности и транзитивности.*

Частично упорядоченному множеству поставлен в соответствие *граф частичного порядка* — невзвешенный ориентированный граф, в котором вершины являются объектами множества X, а ребро, соединяющее вершины a и b, указывает выполнение бинарного отношения z.

Определение 6. *Матрицей предпочтений (матрицей частичного порядка)* $\mathbf{Z}$ *для конечного частично упорядоченного множества $X = \{x_1, ..., x_m\}$ называется матрица инцидентности графа, соответствующего частичному порядку, определенному на X:*

$$\mathbf{Z}(i, k) = z(x_i, x_k). \qquad (1)$$

Пусть задано множество объектов $X = \{x_1, ..., x_m\}$, и на этом множестве задано два различных отношения предпочтения, выраженных отношениями частичного порядка z_1, z_2.

Лемма 1. *Для коэффициента ранговой корреляции Спирмена между z_1 и z_2 выполнено следующее соотношение:*

$$\rho_s \propto -\sum_{i=1}^{m}\left(\sum_{k=1}^{m}(\mathbf{Z}_1(i,k) - \mathbf{Z}_2(i,k))\right)^2, \tag{2}$$

где $\mathbf{Z}_1, \mathbf{Z}_2$ — матрицы предпочтений, соответствующих z_1, z_2.

Лемма 2. *Для коэффициента ранговой корреляции Кендалла между z_1 и z_2 выполнено следующее соотношение:*

$$\tau \propto -\sum_{i=1}^{m}\sum_{k=1}^{m}(\mathbf{Z}_1(i,k) - \mathbf{Z}_2(i,k))^2 = -\|\mathbf{Z}_1 - \mathbf{Z}_2\|_F, \tag{3}$$

где $\|\cdot\|_F$ — норма Фробениуса.

Определение 7. *Полиэдральный конус в $\mathbb{R}^m$ — это множество $\mathcal{X}_0$, такое что*

$$\mathcal{X}_0 = \{\boldsymbol{\chi} \mid \mathbf{A}\boldsymbol{\chi} \le \mathbf{0},\ \boldsymbol{\chi} \in \mathbb{R}^m\}, \tag{4}$$

для некоторой прямоугольной матрицы $\mathbf{A}$.

Пусть на множестве объектов $X = \{x_1, ..., x_m\}$ задано отношение частичного порядка $z: \ z(x_i, x_k) = [x_i \succeq x_k]$. Введем определение конуса частичного порядка, соответствующего множеству X с заданным на нем отношением z.

Определение 8. *Конусом предпочтений $\mathcal{X}_0$ для множества X с определенным на нем отношением частичного порядка $z: \ z = [x_i \succeq x_k]$ называется конус в пространстве $\mathbb{R}^m_+$, сохраняющий порядок относительно z:*

$$\mathcal{X}_0 = \{\boldsymbol{\chi} \in \mathbb{R}^m_+ | \quad x_i \succeq x_k \to \chi_i \ge \chi_k, \quad i,k = 1, ..., m\}. \tag{5}$$

Рассмотрим полиэдральный конус $\mathcal{X}$, представленный неотрицательной комбинацией столбцов матрицы частичного порядка $\mathbf{Z}$, соответствующей множеству X:

$$\mathcal{X} = \{\mathbf{Z}\boldsymbol{\lambda} \mid \boldsymbol{\lambda} \ge \mathbf{0}\}. \tag{6}$$

Теорема 1. *Для конусов* $\mathcal{X}_0$ *и* $\mathcal{X}$*, заданных формулами* (5) *и* (6)*, выполнено* $\mathcal{X} \subset \mathcal{X}_0$.

Глава 2 посвящена проблеме построения модели предпочтений объектов, заданных порядковым описанием. Сформулирована общая постановка задачи с необходимыми ограничениями на монотонность и вид отображения. Поставлена задача максимизации обобщенной ранговой корреляции. Определена линейная полиэдральная модель, использующая в качестве множества допустимых значений сумму Минковского конусов, соответствующих порядковым признакам. Предложены методы оценки параметров линейной полиэдральной модели.

Задано множество X, состоящее из m объектов, $X = \{x_1, ..., x_m\}$. На множестве X задано n отношений предпочтения $z_1, ..., z_n$, описывающих отношения доминирования между каждой парой объектов: $z_j(x_i, x_k) = [x_i \succeq_j x_k]$. В общем случае, каждое из отношений z_j является отношением частичного порядка. Задано целевое отношение предпочтения $z_0: \;\; z_0(x_i, x_k) = [x_i \succeq_0 x_k]$. Требуется построить отображение $f : X \to \mathbb{R}$, сохраняющее порядок на парах объектов:

$$f(x_i) \geq f(x_k), \text{ если } x_i \succeq_0 x_k. \tag{7}$$

От отображения f требуется выполнение условия монотонности по всем отношениям $z_1, ..., z_n$:

$$f(x_i) \geq f(x_k), \text{ если } z_0(\mathbf{x}_i, \mathbf{x}_k) = 1, \tag{8}$$

где отношение доминирования на объектах $x_i \succeq x_k$ выполняется, когда каждая из компонент объекта x_i доминируется соответствующей компонентой объекта x_k, $z_j(\mathbf{x}_i, \mathbf{x}_k) = 1, \; j = 1, ..., n$.

Определение 9. *Конусом целевого предпочтения* $\mathcal{Y}_0$ *для множества* X *с определенным на нем целевым отношением предпочтения* z_0*,*

$$z_0 = [x_i \succeq_0 x_k],$$

называется конус, сохраняющий порядок относительно z_0*:*

$$\mathcal{Y}_0 = \{\boldsymbol{\chi} \in \mathbb{R}^m_+ | \quad x_i \succeq_0 x_k \to \chi_i \geq \chi_k \quad \forall \; i, k = 1, ..., m\}.$$

Будем строить отображение f из множества X во множество действительных чисел $\mathbb{R}$. Обозначим вектор значений функции f на множестве X за $\mathbf{f} \in \mathbb{R}^m$.

Определение 10. *Функцией потерь* $S(f, X, z_0)$ *для задачи восстановления предпочтений* (7) *называется расстояние от вектора* $\mathbf{f}$ *до конуса* $\mathcal{Y}_0$*:*

$$S(f, X, z_0) = \min_{\mathbf{y} \in \mathcal{Y}_0} \|\mathbf{f} - \mathbf{y}\|_2,$$

где $\|\cdot\|_2$ *— евклидова норма.*

Определение 11. *Решением задачи восстановления предпочтений* (7) *называется вектор* $\hat{\mathbf{f}}$*, минимизирующий значений функции потерь* $S(f, X, z_0)$*:*

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f} \in \mathcal{F}} S(f, X, z_0). \tag{9}$$

Конус целевого предпочтения $\mathcal{Y}_0$, как и любой полиэдральный конус, допускает два представления: полиэдральное (4) и задаваемое конечным набором порождающих элементов. Порождающее представление определяется конусом $\mathcal{Y}$, который является подмножеством конуса $\mathcal{Y}_0$, и задается разложением (6):

$$\mathcal{Y}_0 \supset \mathcal{Y} = \{\mathbf{Z}_0 \boldsymbol{\lambda}_0 \mid \boldsymbol{\lambda}_0 \geq \mathbf{0}\}, \tag{10}$$

где $\mathbf{Z}_0$ — матрица частичного порядка, соответствующая целевому отношению предпочтения z_0, определенному на X. Функция потерь $S(f, X, z_0)$ записывается следующим образом:

$$S(f, X, z_0) = \min_{\boldsymbol{\lambda}_0 \geq \mathbf{0}} \|\mathbf{f} - \mathbf{Z}_0 \boldsymbol{\lambda}_0\|_2. \tag{11}$$

Рассматривается подход на основе использования центральной точки конуса $\mathcal{Y}$, то есть случай $\boldsymbol{\lambda}_0 = \mathbf{1}$. В этом случае, функция потерь выглядит следующим образом:

$$S(f, X, z_0) = \|\mathbf{f} - \mathbf{Z}_0 \mathbf{1}\|_2. \tag{12}$$

Для решения задачи (9) каждому из отношений z_j ставится в соответствие конус частичного порядка $\mathcal{X}_j \subset \mathbb{R}^m$, определяемый формулой (5).

Определение 12. *Суммой Минковского двух подмножеств* $\mathcal{X}_1, \mathcal{X}_2$ *линейного пространства называется множество, состоящее из сумм всевозможных векторов из* $\mathcal{X}_1$ *и* $\mathcal{X}_2$*:*

$$\mathcal{X}_1 \oplus \mathcal{X}_2 = \{\boldsymbol{\chi} \mid \boldsymbol{\chi} = \boldsymbol{\chi}_1 + \boldsymbol{\chi}_2,\ \boldsymbol{\chi}_1 \in \mathcal{X}_1,\ \boldsymbol{\chi}_2 \in \mathcal{X}_2\}.$$

Определение 13. *Множеством значений линейной полиэдральной модели* f *на объектах,* $\mathbf{f}(x_1, ..., x_m)$*, называется сумма Минковского конусов предпочтений* $\mathcal{X}_1, ..., \mathcal{X}_n$*:*

$$\mathbf{f}(x_1, ..., x_m) \in \mathcal{X}_1 \oplus ... \oplus \mathcal{X}_n \subset \mathbb{R}^m. \tag{13}$$

Лемма 3. *Значением линейной полиэдральной модели* f *на объектах* $x_1, ..., x_m$ *является неотрицательная линейная комбинация вида*

$$\mathbf{f}(x_1, ..., x_m) = \sum_{j=1}^{n} \mathbf{Z}_j \boldsymbol{\lambda}_j, \ \boldsymbol{\lambda}_j \geq \mathbf{0}. \tag{14}$$

Определение 14. *Линейной полиэдральной моделью* $f : \mathbb{X} \to \mathbb{R}$ *является отображение вида*

$$f(x) = \sum_{j=1}^{n} \mathbf{z}_j^{\mathsf{T}} \boldsymbol{\lambda}_j, \ \boldsymbol{\lambda}_j \geq \mathbf{0}, \tag{15}$$

где $\mathbf{z}_j$ *— вектор индикаторов доминирования объектом* x *объектов* $x_1, ..., x_m$ *согласно отношению предпочтения* z_j*:*

$$\mathbf{z}_j(k) = [x \succeq_j x_k],$$

где $\mathbf{z}_j(k)$ *—* k*-й элемент вектора* $\mathbf{z}_j$*.*

Рассматривается задача минимизации функции потерь $S(f, X, z_0)$ вида (12), где f является линейной полиэдральной моделью вида (14):

$$\mathbf{f}(x_1, ..., x_m) = \sum_{j=1}^{n} \mathbf{Z}_j \boldsymbol{\lambda}_j, \ \boldsymbol{\lambda}_j \geq \mathbf{0},$$

для определения оптимальных параметров $\boldsymbol{\lambda}_j$.

Оптимальные параметры $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ должны минимизировать функцию потерь (12). С учетом явного вида модели $\mathbf{f}$ задача поиска оптимальных параметров записывается следующим образом:

$$(\hat{\boldsymbol{\lambda}}_1, ..., \hat{\boldsymbol{\lambda}}_n) = \underset{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n \geq \mathbf{0}}{\arg\min} \|\mathbf{Z}_0 \mathbf{1} - \sum_{j=1}^{n} \mathbf{Z}_j \boldsymbol{\lambda}_j\|_2. \tag{16}$$

Обозначим за $\hat{\boldsymbol{\lambda}}_j^t$ оценку j-го вектора параметров, полученную на шаге t. Алгоритм итеративной оценки параметров заключается в следующем.

1. На начальном шаге присвоить всем элементам вектора параметров равные значения $\hat{\lambda}_{jk}^0 = 1/m$. В этом случае каждый вектор $\hat{\boldsymbol{\lambda}}_j^0$ является центральной точкой конуса $\mathcal{X}_j$.

2. На шаге t провести последовательную оценку параметров $\hat{\boldsymbol{\lambda}}_j^t$ для всех $j = 1, ..., n$. Для этого предполагаются фиксированными все векторы $\hat{\boldsymbol{\lambda}}_1^t, ..., \hat{\boldsymbol{\lambda}}_{j-1}^t, \hat{\boldsymbol{\lambda}}_{j+1}^{t-1}, ..., \hat{\boldsymbol{\lambda}}_n^{t-1}$, и вектор $\hat{\boldsymbol{\lambda}}_j^t$ отыскивается путем минимизации функции потерь:

$$\hat{\boldsymbol{\lambda}}_j^t = \arg\min_{\boldsymbol{\lambda}_j \geq \mathbf{0}} \|\mathbf{Z}_0\mathbf{1} - \sum_{j'=1}^{j-1} \mathbf{Z}_{j'}\hat{\boldsymbol{\lambda}}_{j'}^t - \sum_{j'=j+1}^{m} \mathbf{Z}_{j'}\hat{\boldsymbol{\lambda}}_{j'}^{t-1} - \mathbf{Z}_j\boldsymbol{\lambda}_j\|_2. \tag{17}$$

   Каждая из задач (17) является задачей неотрицательной линейной регрессии и решается методом активного множества.

Рассматривается задачи минимизации функции потерь (16) на сокращенном пространстве параметров.

Определение 15. *Центральной точкой полиэдрального конуса $\mathcal{X}$, задаваемого конечным набором порождающих элементов $\boldsymbol{\zeta}_k$,*

$$\mathcal{X} = \{\boldsymbol{\chi} \mid \boldsymbol{\chi} = \sum_{k=1}^{r} \lambda_k \boldsymbol{\zeta}_k, \ \lambda_k \geq 0\},$$

является точка $\overline{\mathbf{x}}$ вида

$$\overline{\mathbf{x}} = \frac{1}{r}\sum_{k=1}^{r} \boldsymbol{\zeta}_k.$$

Теорема 2. *При замене конуса $\mathcal{X}_k^w$ его центральной точкой $\overline{\mathbf{x}}_k^w = \frac{1}{n}\sum\limits_{j=1}^{n} w_j \mathbf{z}_{jk}$ линейная полиэдральная модель (14) принимает следующий вид:*

$$\mathbf{f}(\mathbf{x}_1, ..., \mathbf{x}_m) = \hat{\mathbf{Z}}\boldsymbol{\lambda}, \quad \hat{\mathbf{Z}} = \sum_{j=1}^{n} w_j \mathbf{Z}_j, \tag{18}$$

при ограничениях

$$w_j \geq 0, \ \sum_{k=1}^{m} \lambda_k = 1, \ \boldsymbol{\lambda} \geq \mathbf{0}.$$

Оптимизационная задача (16) с учетом регуляризации (18) представляется следующим образом:

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}) = \arg\min_{\mathbf{w},\boldsymbol{\lambda}} \|\mathbf{Z}_0\mathbf{1} - \left(\sum_{j=1}^{n} w_j\mathbf{Z}_j\right)\boldsymbol{\lambda}\|_2, \qquad (19)$$

$$w_j \geq 0, \ \sum_{k=1}^{m}\lambda_k = 1, \ \boldsymbol{\lambda} \geq \mathbf{0}.$$

Лемма 4. *Выполнено неравенство* $\min\limits_{\boldsymbol{\lambda}\geq\mathbf{0},\ \|\boldsymbol{\lambda}\|_1=1} \|\mathbf{Z}_0\mathbf{1} - m\hat{\mathbf{Z}}\boldsymbol{\lambda}\|_2 \leq \|\mathbf{Z}_0 - \hat{\mathbf{Z}}\|_F$.

Доказанное утверждение позволяет свести задачу (19) к задаче восстановления отношения предпочтения на основе максимизации τ-корреляции.

Предложен двухэтапный алгоритм оценки параметров модели (19). На первом этапе строится оценка матрицы частичного порядка $\hat{\mathbf{Z}}$, являющейся линейной комбинацией матриц $\mathbf{Z}_1, ..., \mathbf{Z}_n$,

$$\hat{\mathbf{Z}} = \sum_{j=1}^{n} w_j\mathbf{Z}_j,$$

путем минимизации функции ошибки

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{m}\sum_{k=1}^{m}\left(\mathbf{Z}_0 - \sum_{j=1}^{n} w_j\mathbf{Z}_j\right)^2.$$

На втором этапе строится оценка параметров $\boldsymbol{\lambda}$ линейной комбинации $\hat{\mathbf{Z}}\boldsymbol{\lambda}$ методом неотрицательной линейной регрессии.

Предложен метод решения задачи (12) на основе построения суммы Минковского (13) конусов $\mathcal{X}_1, ..., \mathcal{X}_n$, задаваемых полиэдральным представлением с прямоугольными матрицами $\mathbf{A}_1, ..., \mathbf{A}_n$. Рассматривается линейная полиэдральная модель (13) восстановления отношения предпочтения:

$$\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus ... \oplus \mathcal{X}_n.$$

Для построения вектора оценок объектов предложен метод построения допустимого множества значений, являющегося суммой Минковского выпуклых многогранников $\mathcal{X}_1, ..., \mathcal{X}_n$.

Алгоритм построения суммы Минковского полиэдральных конусов. Суммой Минковского полиэдральных конусов $\mathcal{X}_1 \oplus \mathcal{X}_2 \oplus ... \oplus \mathcal{X}_n$, заданных матрицами $\mathbf{A}_1, ..., \mathbf{A}_n$, является конус

$$\mathcal{X} = \{\mathbf{x} \mid \mathbf{A}^{(n)}\mathbf{x} \leqslant \mathbf{0}\},$$

задаваемый матрицей

$$\mathbf{A}^{(n)} = \mathbf{V}_{n-1}^{\mathsf{T}}\mathbf{A}^{(n-1)},$$

где матрица $\mathbf{A}^{(n-1)}$ определяет конус, являющийся суммой Минковского конусов $\mathcal{X}_1 \oplus \mathcal{X}_2 \oplus ... \oplus \mathcal{X}_{n-1}$, а матрица $\mathbf{V}_{n-1}$ является частью фундаментальной системы решений для уравнения с матрицей $\binom{-\mathbf{A}^{(n-1)}}{\mathbf{A}_n}$, соответствующей неравенствам матрицы $-\mathbf{A}^{(n-1)}$.

В качестве решения задачи восстановления предпочтения в множестве $\mathcal{X}$ рассматривается точка $\hat{\mathbf{y}}$, ближайшую к $\mathbf{y}_0 = \mathbf{Z}_0\mathbf{1}$:

$$\hat{\mathbf{y}} = P_{\mathcal{X}}(\mathbf{y}_0). \tag{20}$$

Глава 3 посвящена задаче построения оценок объектов с использованием экспертных оценок в линейных и порядковых шкалах. Рассматривается задача согласования экспертных оценок.

Задано множество объектов $X = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, матрица плана $\mathbf{X}$, целевая переменная $\mathbf{y}_0 \in \mathbb{R}^m$ и экспертная оценка весов признаков $\mathbf{w}_0 \in \mathbb{R}^n$ для линейной модели $f(\mathbf{w}, \mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$.

Определение 16. *Экспертные оценки* $\mathbf{y}$ *и* $\mathbf{w}$ *называются согласованными, если для них выполняются следующие условия:*

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \quad \mathbf{w} = \mathbf{X}^{+}\mathbf{y}. \tag{21}$$

Обозначим за $\mathbf{y}_0' = \mathbf{X}\mathbf{X}^{+}\mathbf{y}_0$ вектор, являющийся проекцией вектора экспертных оценок $\mathbf{y}_0$ на пространство столбцов матрицы $\mathbf{X}$.

Для разрешения противоречия экспертных данных используются компромиссные оценки:

$$\mathbf{w}_\alpha \in [\mathbf{w}_0, \mathbf{w}_1] \text{ и } \mathbf{y}_\alpha \in [\mathbf{y}_1, \mathbf{y}_0']. \tag{22}$$

Пара векторов $\mathbf{w}_\alpha, \mathbf{y}_\alpha$ для заданного α определяется выражениями

$$\mathbf{w}_\alpha = \alpha\mathbf{w}_0 + (1-\alpha)\mathbf{X}^{+}\mathbf{y}_0', \quad \mathbf{y}_\alpha = (1-\alpha)\mathbf{y}_0' + \alpha\mathbf{X}\mathbf{w}_0.$$

Теорема 3. *Для заданных таким образом векторов* $\mathbf{w}_\alpha, \mathbf{y}_\alpha$ *выполняются условия согласованности* (21).

При отказе от ограничений (22) и отыскании компромиссных оценок в окрестностях векторов $\mathbf{w}_0, \mathbf{y}'_0$ решается задача оптимизации с регуляризующим параметром $\gamma^2 \in [0, +\infty)$:

$$\mathbf{w}_\gamma = \arg\min_{\mathbf{w}\in\mathbb{R}^n} (\varepsilon^2 + \gamma^2\delta^2),$$

где $\varepsilon^2 = \|\mathbf{w}_0 - \mathbf{w}_\gamma\|^2$ and $\delta^2 = \|\mathbf{y}'_0 - \mathbf{y}_\gamma\|^2$. Решением этой задачи являются уточненные экспертные оценки весов признаков

$$\mathbf{w}_\gamma = (\mathbf{X}^\mathsf{T}\mathbf{X} + \gamma^2 I_n)^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y}'_0 + \gamma^2\mathbf{w}_0). \tag{23}$$

Рассматривается случай порядковых экспертных оценок $\mathbf{y}_0, \mathbf{w}_0$. Задано множество объектов $X = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ и матрица плана $\mathbf{X}$. На множестве X определено целевое отношение предпочтения z_0, задаваемое вектором $\mathbf{y}_0 \in Y^m$: $z_0(\mathbf{x}_i, \mathbf{x}_k) = [y_{0i} \geq y_{0k}]$. Кроме того, задано отношение предпочтения на множестве признаков, выражающее степени важности критериев. Это отношение предпочтения задано в виде систем линейных неравенств на вектор весов признаков $\mathbf{w}$ линейной модели $f(\mathbf{w}, \mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$: $w_j \geq w_t$.

Определено два конуса предпочтений $\mathcal{Y}_0$ и $\mathcal{W}_0$, выражающие 1) целевое предпочтение $\mathbf{y}_0$ и 2) важность критериев. Рассматривается полиэдральное представление этих конусов с матрицами $\mathbf{A}^m$ и $\mathbf{A}^n$, соответственно:

$$\mathcal{Y}_0 = \{\mathbf{x}|\mathbf{A}^m\mathbf{x} \leqslant 0\}, \quad \mathcal{W}_0 = \{\mathbf{w}|\mathbf{A}^n\mathbf{w} \leqslant 0\}.$$

Порядковые экспертные оценки называются *согласованными*, если конусы $\mathbf{X}\mathcal{W}_0$ и $\mathcal{Y}'_0$ имеют непустое пересечение, отличное от точки $\mathbf{0}$. Если конусы $\mathbf{X}\mathcal{W}_0$ и $\mathcal{Y}'_0$ в качестве пересечения содержат только точку $\mathbf{0}$, порядковые экспертные оценки называются *несогласованными*.

Согласованная пара оценок $\hat{\mathbf{w}}, \hat{\mathbf{y}}$ должна удовлетворять следующим ограничениям:

$$\begin{cases} \mathbf{A}^n\hat{\mathbf{w}} \leqslant \mathbf{0}, \\ \hat{\mathbf{y}} = \mathbf{X}\mathbf{X}^+\mathbf{y}, \\ \mathbf{A}^m\mathbf{y} \leqslant \mathbf{0}. \end{cases}$$

Требуется найти в конусах $\mathcal{W}_0$ и $\mathcal{Y}_0$ векторы $\hat{\mathbf{w}}$ и $\mathbf{y}_1$, такие что

$$(\hat{\mathbf{w}}, \mathbf{y}_1) = \underset{\mathbf{w}\in\mathcal{W}_0, \mathbf{y}\in\mathcal{Y}_0}{\arg\min} \|\mathbf{X}^{+}\mathbf{y} - \mathbf{w}\|_2, \quad \|\mathbf{X}^{+}\mathbf{y}\|_2 = 1, \|\mathbf{w}\|_2 = 1, \tag{24}$$

Векторы $\mathbf{a} = \mathbf{y}^{(2k)}$ и $\mathbf{b} = \mathbf{w}^{(2k+1)}$ являются решениями двух последовательно решаемых оптимизационных задач, полагая вектор $\mathbf{w}^{(0)} = \mathbf{w}_0$ на шаге $k = 0$, где $\mathbf{w}_0$ — центральная точка конуса $\mathcal{W}_0$.

Задача $2k$:		Задача $2k+1$:	
minimize	$\|\mathbf{X}^{+}\mathbf{a} - \mathbf{w}^{(2k)}\|_2$	minimize	$\|\mathbf{X}^{+}\mathbf{y}^{(2k+1)} - \mathbf{b}\|_2$
subject to	$(\mathbf{X}^{+}\mathbf{a})^{\mathsf{T}}\mathbf{X}^{+}\mathbf{a} = 1,$	subject to	$\mathbf{b}^{\mathsf{T}}\mathbf{b} = 1,$
	$\mathbf{A}^m\mathbf{a} \leqslant 0.$		$\mathbf{A}^n\mathbf{b} \leqslant 0.$

С использованием порождающего представления конусов для решения задачи согласования порядковых оценок предложен итеративный алгоритм, основанный на последовательном решении задач неотрицательной линейной регрессии:

Задача $2k$:	Задача $2k+1$:
$\hat{\boldsymbol{\lambda}}_0 = \min\limits_{\boldsymbol{\lambda}_0 \geq \mathbf{0},\ \|\boldsymbol{\lambda}_0\|=1} \|\mathbf{Z}_0\boldsymbol{\lambda}_0 - \mathbf{X}_w\hat{\boldsymbol{\lambda}}_w^{(2k-1)}\|_2^2$	$\hat{\boldsymbol{\lambda}}_w = \min\limits_{\boldsymbol{\lambda}_w \geq \mathbf{0},\ \|\boldsymbol{\lambda}_w\|=1} \|\mathbf{Z}_0\hat{\boldsymbol{\lambda}}_0^{(2k)} - \mathbf{X}_w\boldsymbol{\lambda}_w\|_2^2$

В **главе 4** приводится описание разработанного программного комплекса и результаты анализа методов.

В качестве прикладных задач рассматривается задача категоризации редких видов Красной книги РФ, задача ранжирования заповедников и задача порядковой классификации на наборах данных из репозитория UCI: Pyrimidines, Machine CPU, Housing, Computer Activity, Abalone и Car. Перечисленные наборы данных, кроме последнего, относятся к задаче регрессии, поэтому была проведена дополнительная процедура дискретизациии целевой переменной на пять уровней, содержащих равное число объектов, для решения задачи пятиклассовой порядковой классификации.

Каждый набор данных был разбит случайным образом на обучающую и контрольную части. Это разбиение было проведено независимо 100 раз. Для оценки качества были использована средняя абсолютная потеря и средняя потеря Хэмминга. Для сравнения методов на данных порядковой природы была произведена порядковая трансформация признаков исходных наборов данных.

Таблица 1. Результаты на данных UCI: порядковые признаки

	Средняя абсолютная потеря (±0.01)					Средняя потеря Хэмминга (±0.01)				
Набор данных	SVM	POF	Trees	OW	KNN	SVM	POF	Trees	OW	KNN
Pyrimidines	0.57	0.58	0.60	0.62	**0.49**	**0.71**	0.77	0.79	0.79	0.76
MachineCPU	0.51	**0.39**	0.47	**0.40**	0.43	0.65	**0.45**	0.56	0.47	0.51
Boston	**0.40**	0.48	**0.40**	0.43	**0.41**	0.49	0.68	**0.46**	0.50	0.51
Computer	0.44	0.69	0.41	**0.37**	0.45	0.53	1.38	**0.45**	**0.44**	0.55
Abalone	0.78	**0.59**	**0.57**	**0.58**	**0.59**	1.78	0.92	**0.76**	0.85	0.89
Cars	0.19	0.19	0.08	0.16	**0.06**	0.24	0.26	**0.08**	0.19	**0.07**
RedBook	0.56	0.61	**0.50**	**0.49**	0.59	0.66	0.74	**0.55**	0.59	0.72

Также, как и в случае с целевой переменной, была произведена дискретизация каждого признака на пять уровней, содержащих одинаковое количество объектов.

Было проведено сравнение методов с двумя стандартными алгоритмами порядковой классификации и методом, учитывающим монотонные ограничения на признаки.

Результаты для преобразованных наборов данных показаны в таблице 1. Предложенный метод оценки матрицы предпочтений (OW) продемонстрировал лучший результат в терминах средней абсолютной потери.

Задача категоризации Красной книги заключается в следующем: каждому объекту Красной книги необходимо поставить в соответствие категорию (вымершие, вымершие в природе, находящиеся в критическом состоянии и пр.), отражающую степень угрожаемости объекту. Такая категоризация является монотонной. В качестве признакового описания задано экспертное описание объектов.

Результаты тестирования для задачи категоризации Красной книги показаны в последней строке таблицы 1. В отличие от остальных наборов данных, для этой задачи была использована схема разбиения leave-one-out. Предложенный метод продемонстрировал наилучший результат наряду с методом Trees.

В заключении представлены **основные результаты** диссертационной работы.

1. Рассмотрены различные задачи обучения по предпочтениям. Для описания отношения предпочтения предложено использовать граф, соответствующий частично упорядоченному множеству, и конус предпочтений. Рассмотрено два представления конуса предпочтений: полиэдральное и

порождающее. Установлено, что столбцы матрицы предпочтений являются порождающими элементами конуса предпочтений. Кроме того, с использованием матрицы предпочтений получены обобщенные формулы для известных порядковых метрик, таких как τ-корреляция и площадь под кривой.

2. Предложено несколько подходов для решения задачи восстановления отношения предпочтения на множестве объектов, заданных порядковым описанием. Предложен подход на основе восстановления матрицы предпочтений, основанный на максимизации обобщенной τ-корреляции. Разработан метод на основе построения суммы Минковского конусов, соответствующих порядковым признакам. Предложен ряд методов оценки параметров построенной суперпозиции. Установлено соответствие использования центральных точек конусов с задачей восстановления матрицы предпочтений.

3. Разработаны методы согласования экспертных оценок. Введено понятие согласованности оценок. Рассмотрено два типа оценок: линейные и порядковые. В линейном случае предложена процедура согласования оценок с использованием структурного параметра. В порядковом случае предложен метод согласования на основе минимизации расстояния между векторами в конусах, соответствующих порядковым экспертным оценкам.

4. Разработан программный комплекс методов построения модели предпочтений. Приведено решение задачи категоризации редких видов из Красной книги РФ, а также задачи ранжирования заповедников. Проведено сравнение разработанных методов для решения задачи порядковой классификации на стандартных наборах данных из репозитория UCI. Разработанные методы показывают улучшение качества в сравнении с известными подходами.

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК.

1. Kuznetsov M.P., Strijov V.V. Methods of expert estimations concordance for integral quality estimation // Expert Systems with Applications. — 2014. — V. 41, no. 4. — P. 1988-1996.

2. M. Kuznetsov, M. Clausel, M.-R. Amini, E. Gaussier and V. Strijov. Supervised topic classification for modeling a hierarchical conference structure // Neural Information Processing. Springer International Publishing — 2015. — P. 90-97.

3. Stenina M.M., Kuznetsov M.P, Strijov V.V. Ordinal classification using pareto fronts // Expert Systems with Applications. — 2015. — V. 42, no. 14. — P. 5947–5953.

4. Кузнецов М. П., Стрижов В. В., Медведникова М.М. Алгоритм многоклассовой классификации объектов, описанных в ранговых шкалах // Научно-технический вестник СПб ГПУ. Информатика. Телекоммуникации. Управление. — 2012. — Т. 5. — С. 92–95.

5. Стрижов В. В., Кузнецов М. П., Рудаков К. В. Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Математическая биология и биоинформатика. — 2012. — Т. 7(1). — С. 345–359.

6. Медведникова М. М., Стрижов В. В., Кузнецов М. П. Алгоритм многоклассовой монотонной парето-классификации с выбором признаков // Известия Тульского государственного университета, Естественные науки. — 2012. — Т. 3. — С. 132–141.

Остальные публикации.

7. Кузнецов М.П., Ивкин Н.П. Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных. — 2015. — Т. 1, № 11. — С. 1471-1483.

8. Кузнецов М.П. Комбинирование отношений порядка для восстановления предпочтения на наборе объектов // Математические методы распознавания образов ММРО-15. Тезисы докладов 17-й Всероссийской конференции с международным участием. г. Светлогорск, Калининградская обл. М.: Торус пресс. — 2015. — С. 18-19.

9. Relevance-aware filtering of tuples sorted by an attribute value via direct optimization of search quality metrics / Nikita V Spirin, Mikhail Kuznetsov, Julia Kiseleva et al. // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval / ACM. — 2015. — P. 979–982.

10. A. M. Katrutsa, M. P. Kuznetsov, V. V. Strijov, K. V. Rudakov Metric concentration search procedure using reduced matrix of pairwise distances // Intelligent Data Analysis. — 2015. — V. 19, no. 5. — P. 1091–1108.

11. Кузнецов М.П. Построение интегрального индикатора в ранговых шкалах с использованием копул для анализа совместного распределения критериев // Машинное обучение и анализ данных. — 2012. — Т. 1. — С. 411–419.

12. Кузнецов М.П., Стрижов В.В. Построение интегрального индикатора с использованием ранговой матрицы описаний // Интеллектуализация обработки информации. Доклады 9-й международной конференции. М.: Торус пресс. — 2012. — С. 130–132.

13. Кузнецов М.П. Уточнение ранговых экспертных оценок с использованием монотонной интерполяции // Машинное обучение и анализ данных. — 2011. — Т. 1, № 2. — С. 154–162.

Кузнецов Михаил Павлович

ПОСТРОЕНИЕ МОДЕЛЕЙ ОБУЧЕНИЯ ПО ПРЕДПОЧТЕНИЯМ С ИСПОЛЬЗОВАНИЕМ ПОРЯДКОВЫХ ЭКСПЕРТНЫХ ОЦЕНОК

АВТОРЕФЕРАТ

Подписано в печать ___ . ___ . 2016. Формат 60 × 84 1/16. Усл. печ. л. 1,0.
Тираж 100 экз. Заказ № ______ .

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»

Отдел оперативной полиграфии «Физтех-полиграф» 141700, Московская обл., г. Долгопрудный, Институтский пер., 9