На правах рукописи

Ткачев Юрий Игоревич

МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТИ КОЛЛЕКТИВАМИ РАСПОЗНАЮЩИХ АЛГОРИТМОВ

Специальность 05.13.17 —
Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2013

Работа выполнена в Федеральном государственном бюджетном учреждение науки Вычислительный центр им. А.А. Дородницына Российской академии наук.

Научный руководитель:	доктор физико-математических наук, профессор Рязанов Владимир Васильевич.
Официальные оппоненты:	доктор физико-математических наук Шибзухов Заур Мухадинович, ведущий научный сотрудник Научно-аналитического управления, Всероссийский центр мониторинга и прогнозирования чрезвычайных ситуаций природного и природно-техногенного характера, Министерство по делам гражданской обороны и чрезвычайных ситуаций РФ.
	кандидат физико-математических наук Челноков Федор Борисович, старший технический менеджер, Московский филиал Корпорации Алаин Текноложди Ресерч энд Девелопмент, Инк. (США).
Ведущая организация:	Федеральное государственное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)».

Защита состоится «24» октября 2013 г. на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждение науки Вычислительный центр им. А.А.Дородницына Российской академии наук по адресу: 119333, г. Москва, ул. Вавилова, д. 40. С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан «17» сентября 2013 г.

Ученый секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор В.В.Рязанов

Общая характеристика работы

Диссертационная работа посвящена проблеме восстановления зависимостей по выборкам прецедентов. Предлагаются модели «байесовский корректор» и «линейный корректор» для решения данной задачи, основанные на решении набора специальных задач распознавания, построенных на исходной обучающей выборке, и последующей коррекции в пространстве значений целевого признака. Исследуются свойства корректности и устойчивости рассматриваемых моделей.

Актуальность темы. В настоящее время существуют различные параметрические и непараметрические подходы к восстановлению зависимостей по выборкам прецедентов: линейная, полиномиальная и обобщенная линейная модели, метод опорных векторов, логистическая регрессия, нейросетевые алгоритмы, ядерное сглаживание, регрессионные деревья, случайный лес, регрессия на основе классификации. Следует отметить существенные ограничения существующих подходов при решении реальных задач. Параметрические подходы требуют априорного знания аналитического вида функций. Наличие разнотипных признаков требует привлечения дополнительных средств описания объектов в единой шкале. Непараметрические подходы используют, как правило, методы частотной оценки в некоторой окрестности, при этом возникают проблемы выбора окрестности и функций расстояния, учета фактора различной важности признаков и т.п.

В то же время, случай дискретной величины (стандартная задача распознавания) в настоящее время достаточно хорошо изучен. Более 20 лет тому назад Ю.И.Журавлевым было отмечено, что задача распознавания может рассматриваться как задача интерполяции специальных функций, когда независимые переменные (значения признаков) могут быть фактически произвольны, а зависимая величина принимает конечное число значений. В настоящее время существуют различные модели и конкретные алгоритмы для решения задач распознавания, для которых исследованы свойства корректности и устойчивости (алгебраический подход, логические корректоры и др.).

Актуальной задачей является разработка методов восстановления зависимостей по выборкам прецедентов, основанных на решении набора специ-

альных задач распознавания и последующей коррекции в пространстве значений целевого признака. При этом основные трудности, связанные со сравнением объектов в признаковом пространстве (разнотипность и различная информативность признаков, согласование метрик для отдельных признаков, и др.), переносятся на уровень решения задач распознавания.

Целью диссертационной работы является разработка и исследование методов решения задачи восстановления зависимостей по выборкам прецедентов, основанных на решении набора специальных задач распознавания, построенных на исходной обучающей выборке, и последующей коррекции в пространстве значений целевого признака.

Основные положения, выносимые на защиту.

1. Общий подход к формированию задач распознавания и вычисления значения зависимой величины как коллективного решения.

2. Модели восстановления зависимостей «байесовский корректор» и «линейный корректор».

3. Доказательства свойств корректности и устойчивости моделей восстановления зависимостей «байесовский корректор» и «линейный корректор».

4. Результаты практической апробации моделей восстановления зависимостей на реальных прикладных задачах.

Научная новизна. Автором разработаны методы решения задачи восстановления зависимостей по выборкам прецедентов, основанных на решении набора специальных задач распознавания, построенных на исходной обучающей выборке, и последующей коррекции в пространстве значений целевого признака. Доказаны теоретические утверждения о свойствах корректности и устойчивости предложенных моделей восстановления зависимостей.

Теоретическая значимость. В работе предложен новый подход к решению задач восстановления зависимостей по выборкам прецедентов, приводятся доказательства свойств корректности и устойчивости предложенных методов.

Практическая значимость. Разработанные методы восстановления зависимостей по выборкам прецедентов применимы к реальным прикладным задачам, что подтверждается практической апробацией.

Достоверность результатов подтверждена строгими математическими доказательствами теоретических утверждений.

Апробация работы. Основные результаты работы докладывались на следующих научных конференциях:

- 14-я всероссийская конференция «Математические методы распознавания образов» (Суздаль, 2009 год);
- 2-я международная конференция «Классификация, прогнозирование, анализ данных» CFDM-2010 (Варна, Болгария, 2010 год);
- 10-я международная конференция «Распознавание образов и анализ изображений: новые информационные технологии – РОАИ-10-2010» (Санкт-Петербург, 2010 год).

Публикации. Материалы диссертации опубликованы в 4 научных статьях [1–4], из них 2 работы [2, 3] — в журналах, включенных в Перечень ведущих рецензируемых научных журналов и изданий.

Структура и объем диссертации. Работа состоит из оглавления, введения, трех глав, заключения и списка литературы. Содержание работы изложено на 47 страницах. Список литературы включает 49 наименований. Текст работы иллюстрируется 1 таблицей.

Содержание работы

Во введении обоснована актуальность диссертационной работы, сформулированы цели и задачи, аргументирована научная значимость исследования, представлены результаты и положения, выносимые на защиту, приведена краткая структура диссертации.

В первой главе описывается общий подход к решению задачи восстановления зависимостей по выборкам прецедентов коллективами распознающих алгоритмов. Рассматриваются методы формирования разбиений вещественной оси на интервалы. Описывается алгоритм формирования набора

специальных задач распознавания, построенных на исходной обучающей выборке. Рассматриваются модели восстановления зависимостей по выборкам прецедентов «байесовский корректор» и «линейный корректор». Исследуются свойства корректности (модель не делает ошибок на обучающей выборке) и квазикорректности относительно разбиения области значений целевого признака (суммарная ошибка модели на обучающей выборке не превосходит заданной величины) рассматриваемых моделей.

Имеется выборка $\{(\vec{x}_i, y_i)\}_{i=1}^m$, $\vec{x} = (x_1, \ldots, x_d)$ — описание объекта, $x_j \in \mathbf{M}_j$, $j = 1, \ldots, d$, $\mathbf{M}_j$ — множество произвольной природы; $y \in R$ — значение целевого признака, $R \subset \mathbb{R}$. Обозначим m' — число различных значений целевого признака в обучающей выборке, $m' = |\{y_i\}_{i=1}^m|$.

Рассмотрим разбиение области значений целевого признака $\mathbf{\Delta} = \{\Delta_1, \ldots, \Delta_n\}$ на интервалы $\Delta_1 = (d_0, d_1], \ldots, \Delta_n = (d_{n-1}, d_n]$. Без ограничения общности считаем, что объекты обучающей выборки упорядочены по возрастанию значения целевого признака, тогда $y_1, \ldots, y_{m_1} \in \Delta_1, y_{m_1+1}, \ldots, y_{m_2} \in \Delta_2, \ldots, y_{m_{n-1}+1}, \ldots, y_{m_n} \in \Delta_n$.

Возьмем число L, $2 \leqslant L \leqslant n$, и определим N разбиений отрезка R на L интервалов, поставив каждому разбиению в соответствие вектор с целочисленными компонентами $\mathbf{k}_i = (0, k_i^{(1)}, \ldots, k_i^{(L-1)}, n), i = 1, \ldots, N$, $k_i^{(j)} < k_i^{(j+1)} < n$. Каждое разбиение отрезка R определяет разбиение множества $\mathbf{M} = \mathbf{M}_1 \times \cdots \times \mathbf{M}_d$ на L непересекающихся подмножеств $K_1^i, \ldots, K_L^i$ (классов): $\mathbf{M} = \bigcup\limits_{t=1}^{L} K_t^i$, $\nu \neq \mu \Rightarrow K_\nu^i \cap K_\mu^i = \varnothing$ согласно правилу: объект $\vec{x}$ принадлежит классу K_j^i разбиения $\mathbf{K}^i = \{K_1^i, \ldots, K_L^i\}$ тогда и только тогда, когда $y = f(\vec{x}) \in \Delta_k$ и $k \in (k_{i,j}, k_{i,j+1}]$. Каждое разбиение $\mathbf{K}^i$ определяет стандартную задачу распознавания Z_i с L классами. Пусть A_i — некоторый алгоритм решения задачи распознавания Z_i, относящий произвольный объект $\vec{x}$ к одному из классов $K_{a_i}^i$, $a_i = 1, \ldots, L$.

Общий подход к решению задачи восстановления зависимостей коллективами распознающих алгоритмов состоит из подзадач:

1. формирование разбиений вещественной оси на интервалы Δ_k;
2. формирование обучающих выборок задач распознавания Z_i, выбор классификаторов A_i, их обучение;
3. вычисление значения зависимой величины на основе коллективного решения классификаторов.

В разделе 1.1 рассматривается задача формирования разбиения вещественной оси на фиксированное число интервалов n, приводится алгоритм решения данной задачи методом динамического программирования.

Для формирования разбиения вещественной оси на интервалы Δ_k требуется определить номера граничных объектов m_k: $y_{m_{k-1}} < y_i \leqslant y_{m_k}$. Обозначим $c_k = \frac{\sum_{i=m_{k-1}+1}^{m_k} y_i}{m_k - m_{k-1}}$ — выборочное среднее для интервала Δ_k, $k = 1, \ldots, n$.

В задаче динамического программирования используется функционал $Q(\boldsymbol{\Delta}) = \sum\limits_{k=1}^{n} \sum\limits_{i=m_{k-1}+1}^{m_k} (y_i - c_k)^2$, оптимизация проводится по номерам граничных объектов m_k, $k = 1, \ldots, n$.

Построение разбиения области значений целевого признака осуществляется на основе обучающей выборки и не зависит от алгоритмов распознавания, используемых в дальнейшем.

В разделе 1.2 рассматривается модель восстановления зависимости «байесовский корректор», исследуются свойства модели: корректность и квазикорректность относительно разбиения области значений целевого признака.

Считаем декартово произведение $\mathbf{K}^1 \times \cdots \times \mathbf{K}^N \times \boldsymbol{\Delta}$ множеством элементарных событий, событие $(K^1_{a_1}, \ldots, K^N_{a_N}, \Delta_k)$ означает: «объект $\vec{x}$ отнесен алгоритмом A_1 в класс a_1,..., алгоритмом A_N в класс a_N, $y = f(\vec{x}) \in \Delta_k$». Вероятность такого события обозначается как $\mathsf{P}(K^1_{a_1}, \ldots, K^N_{a_N}, \Delta_k)$ или $\mathsf{P}(\vec{x}, \Delta_k)$.

По формуле Байеса имеем

$$\mathsf{P}(\Delta_k|K^1_{a_1},\ldots,K^N_{a_N}) = \frac{\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N},\Delta_k)}{\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N})} = \\ = \frac{\mathsf{P}(\Delta_k)}{\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N})}\,\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N}|\Delta_k) \quad (1)$$

Пусть классификаторы статистически независимы, тогда $\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N}|\Delta_k) = \prod\limits_{i=1}^{N}\mathsf{P}(K^i_{a_i}|\Delta_k)$, $\mathsf{P}(K^1_{a_1},\ldots,K^N_{a_N}) = \prod\limits_{i=1}^{N}\mathsf{P}(K^i_{a_i})$, и формулу Байеса можно записать в виде:

$$\mathsf{P}(\Delta_k|\vec{x}) = \mathsf{P}(\Delta_k|K^1_{a_1},\ldots,K^N_{a_N}) = \frac{\mathsf{P}(\Delta_k)}{\prod\limits_{i=1}^{N}\mathsf{P}(K^i_{a_i})}\prod_{i=1}^{N}\mathsf{P}(K^i_{a_i}|\Delta_k) \quad (2)$$

Определение 1 *Байесовским корректором называется функция F:*

$(\mathsf{P}(\Delta_1|\vec{x}),\ldots,\mathsf{P}(\Delta_n|\vec{x})) \;\to\; \mathbb{R}$.

Определение 2 *«Ответом по среднему» байесовского корректора для объекта $\vec{x}$ называется величина $\tilde{y} = \sum\limits_{k=1}^{n}\mathsf{P}(\Delta_k|\vec{x})c_k$.*

Определение 3 *«Ответом по максимуму» байесовского корректора для объекта $\vec{x}$ называется величина $\hat{y} = c_k$, где $k = \arg\max\limits_{j=1}^{n}\mathsf{P}(\Delta_j|\vec{x})$.*

Алгоритм восстановления зависимости

1. формирование разбиений вещественной оси на интервалы Δ_k, $k = 1,\ldots,n$;
2. задание разбиений $\mathbf{K}^i$, $i = 1,\ldots,N$, формирование обучающих выборок задач распознавания Z_i, $i = 1,\ldots,N$, выбор классификаторов A_i, $i = 1,\ldots,N$, их обучение;
3. вычисление оценок вероятностей $\mathsf{P}(K^i_j|\Delta_k)$, $\mathsf{P}(\Delta_k)$, $\mathsf{P}(K^i_j)$, $i = 1,\ldots,N$, $j = 1,\ldots,L$, $k = 1,\ldots,n$.

Алгоритм вычисления значения зависимой величины

1. классификация алгоритмами $A_1, \ldots, A_N$ объекта $\vec{x}$;
2. вычисление значений условных вероятностей $\mathsf{P}(\Delta_k|\vec{x})$, $k = 1, \ldots, n$ по формулам (2);
3. вычисление «ответа по среднему» $\tilde{y}$ или «ответа по максимуму» $\hat{y}$.

Для вычисления параметров модели восстановления зависимостей ставится следующая оптимизационная задача:

$$F = \sum_{i=1}^{m} (y_i - \tilde{y}_i)^2 \to \min_{\mathsf{P}(K_j^i|\Delta_k),\mathsf{P}(\Delta_k),\mathsf{P}(K_j^i)}$$

при ограничениях:

$$\sum_{j=1}^{L} \mathsf{P}(K_j^i) = 1, i = 1, \ldots, N;$$

$$\sum_{k=1}^{n} \mathsf{P}(\Delta_k) = 1, \mathsf{P}(\Delta_k) > 0, k = 1, \ldots, n;$$

$$\sum_{k=1}^{n} \mathsf{P}(\Delta_k)\,\mathsf{P}(K_j^i|\Delta_k) = \mathsf{P}(K_j^i), \mathsf{P}(K_j^i|\Delta_k) \geqslant 0, i = 1, \ldots, N, j = 1, \ldots, L.$$

Аналитическим решением данной задачи при $L = 2, N = n - 1$ и фиксированных значениях вероятностях интервалов $\mathsf{P}(\Delta_k) = \alpha_k$ являются величины:

$$\begin{aligned}
\mathsf{P}(K_1^i|\Delta_k) &= \begin{cases} 0, & i < k, \\ 1, & i = k, \\ \frac{\alpha_i}{1 - \sum_{t=1}^{i-1} \alpha_t}, & i > k; \end{cases} \\
\mathsf{P}(K_2^i|\Delta_k) &= 1 - \mathsf{P}(K_1^i|\Delta_k); \\
\mathsf{P}(K_1^i) &= \frac{\alpha_i}{1 - \sum_{t=1}^{i-1} \alpha_t}; \\
\mathsf{P}(K_2^i) &= 1 - \mathsf{P}(K_1^i); \\
i = 1, \ldots, n; k &= 1, \ldots, n.
\end{aligned} \tag{3}$$

Определение 4 *Модель обладает свойством* $(*)$, *если для объекта* $\vec{x}_i$, $i = 1, \ldots, m$, *обучающей выборки* $\mathsf{P}(\Delta_k|\vec{x}_i) \neq 0 \Leftrightarrow y_i \in \Delta_k$.

Утверждение 1.1 *Байесовский корректор с оценками вероятностей* (3) *обладает свойством* $(*)$.

Альтернативный подход к вычислению оценок вероятности для модели восстановления зависимости заключается в использовании частотных оценок.

Частотные оценки условных вероятностей

$$\mathsf{P}(K_1^i|\Delta_k) = \begin{cases} 1, & i \geqslant k, \\ 0, & i < k \end{cases}, i = 1, \ldots, n-1, k = 1, \ldots, n;$$

$$\mathsf{P}(K_2^i|\Delta_k) = \begin{cases} 1, & i < k, \\ 0, & i \geqslant k \end{cases}, i = 1, \ldots, n-1, k = 1, \ldots, n.$$

Частотные оценки вероятностей классов $\mathsf{P}(K_1^i) = \sum\limits_{k=1}^{n} \mathsf{P}(K_1^i|\Delta_k)\,\mathsf{P}(\Delta_k) = \sum\limits_{k=1}^{i} \alpha_k$; $\mathsf{P}(K_2^i) = \sum\limits_{k=1}^{n} \mathsf{P}(K_2^i|\Delta_k)\,\mathsf{P}(\Delta_k) = \sum\limits_{k=i+1}^{n} \alpha_k$, $i = 1, \ldots, n-1$.

Утверждение 1.2 *Байесовский корректор с частотными оценками вероятностей и «ответом по среднему» или «ответом по максимуму» обладает свойством* $(*)$.

При практическом построении $\mathsf{P}(K_j^i|\Delta_k)$ возникают ситуации, когда классификаторы относят объект $\vec{x}$ к классам так, что $\mathsf{P}(K_j^i|\Delta_k) = 0$, $k = 1, \ldots, n$. Рассматривается подход, позволяющий бороться с такими ситуациями.

Рассмотрим линейную комбинацию оценок вероятностей

$$\tilde{\mathsf{P}}(K_j^i|\Delta_k) = \frac{\sum\limits_{t=\max\{-d,1-k\}}^{\min\{d,n-k\}} w_t\,\mathsf{P}(K_j^i|\Delta_{k+t})}{\sum\limits_{t=\max\{-d,1-k\}}^{\min\{d,n-k\}} w_t}, i = 1, \ldots, n, j = 1, \ldots, L, k = 1, \ldots, n$$

$$\tilde{\mathsf{P}}(K_j^i) = \sum_{k=1}^{n} \mathsf{P}(\Delta_k)\tilde{\mathsf{P}}(K_j^i|\Delta_k)$$

Величины $\tilde{\mathsf{P}}(K_j^i|\Delta_k)$ формально являются вероятностями.

Назовем величину d шириной окна сглаживания, величины $w_{-d}, \ldots, w_d$ — весами сглаживания. Процесс замены $\mathsf{P}(K_j^i|\Delta_k) \to \tilde{\mathsf{P}}(K_j^i|\Delta_k)$ назовем сглаживанием.

На веса сглаживания накладываются ограничения:

- неотрицательность: $w_t \geqslant 0, t = -d, \ldots, d$;
- симметричность: $w_t = w_{-t}, t = 1, \ldots, d$;
- характер убывания: $w_t = u^{d-|t|}, u > 1, t = -d, \ldots, d$.

Определение 5 *Модель обладает свойством* $(**)$, *если для объекта* $\vec{x}_i, i = 1, \ldots, m$ *обучающей выборки:* $y_i \in \Delta_k \Rightarrow \mathsf{P}(\Delta_k|\vec{x}_i) > \mathsf{P}(\Delta_{k_1}|\vec{x}_i), k \neq k_1$.

Утверждение 1.3 *Байесовский корректор со сглаженными частотными оценками вероятностей и «ответом по максимуму» обладает свойством* $(**)$.

Определение 6 *Модель восстановления зависимости называется* ***квазикорректной*** *относительно разбиения* $\boldsymbol{\Delta}$, *если для обучающей выборки* $\{(y_i, \vec{x}_i)\}_{i=1}^m$ *выполнено* $y_i \in \Delta_k \Rightarrow \hat{y}_i \in \Delta_k$, $i = 1, \ldots, m$.

Определение 7 *Модель восстановления зависимости называется* ***корректной***, *если для обучающей выборки* $\{(y_i, \vec{x}_i)\}_{i=1}^m$ *выполнено* $\tilde{y}_i = y_i, i = 1, \ldots, m$.

Определение 8 *Моделью* $\mathbb{A}_1$ *называется байесовский корректор над корректными классификаторами с аналитическими оценками вероятностей при использовании «ответа по среднему».*

Определение 9 *Моделью* $\mathbb{A}_2$ *называется байесовский корректор над корректными классификаторами с частотными оценками вероятностей при использовании «ответа по максимуму».*

Определение 10 *Моделью* $\mathbb{A}_2^d$ *называется байесовский корректор над корректными классификаторами со сглаженными частотными оценками вероятностей и шириной окна сглаживания* d *при использовании «ответа по максимуму».*

Теорема 1.1 *Модели* $\mathbb{A}_1, \mathbb{A}_2$ *и* $\mathbb{A}_2^d$ *при непротиворечивой обучающей информации являются квазикорректными относительно разбиения* $\boldsymbol{\Delta}$.

Теорема 1.2 *Модели* $\mathbb{A}_1, \mathbb{A}_2$ *и* $\mathbb{A}_2^d$ *при непротиворечивой обучающей информации и разбиении* $\mathbf{\Delta} : n = m'$ *области значений целевого признака являются корректными.*

Для обучения корректной модели восстановления зависимости с разбиением области значений целевого признака на m' интервалов необходимо обучить $m' - 1$ классификаторов, т.е. число классификаторов сравнимо с количеством объектов обучающей выборки. Рассмотрим свойства байесовского корректора для разбиения $\mathbf{\Delta} : |\mathbf{\Delta}| = n, n < m'$.

Погрешность модели восстановления зависимости есть $\sum\limits_{i=1}^{m} |y_i - c_{k_i}|^2$, $y_i \in \Delta_{k_i}$, $i = 1, \ldots, m$. Данное выражение совпадает с погрешностью разбиения $Q(\mathbf{\Delta})$, т.е. для построения модели с заданной погрешностью δ необходимо построить разбиение с погрешностью δ.

Утверждение 1.4 *Алгоритмическая сложность построения разбиения для байесовского корректора, дающего суммарную ошибку* $Q(\mathbf{\Delta}) < \delta$, *равна* $O(m^3 \log m)$.

В разделе 1.3 рассматривается модель восстановления зависимости «линейный корректор», исследуются свойства модели: корректность и квазикорректность относительно разбиения области значений целевого признака.

Обозначим $u_i^{(j)} = \frac{\sum_{k=k_{i,j}+1}^{k_{i,j+1}} c_k}{k_{i,j+1} - k_{i,j}}, i = 1, \ldots, N, j = 1, \ldots, L$.

Составим вектор ответов $\vec{u}(\vec{x}) = (1, u_1(\vec{x}), \ldots, u_N(\vec{x}))$, $u_i(\vec{x}) = u_i^{(j)}$, если классификатор A_i отнес объект $\vec{x}$ в класс K_j^i.

Определение 11 *Линейным корректором называется функция* $f : X \to \mathbb{R}$, $f(\vec{x}) = (\vec{u}(\vec{x}), \vec{w})$, $\vec{w}$ *- вектор весов,* $\vec{w} \in \mathbb{R}^{N+1}$.

Вектор весов есть решение оптимизационной задачи $\sum\limits_{i=1}^{m} (y_i - f(\vec{x}_i))^2 \to \min\limits_{\vec{w}}$.

Алгоритм восстановления зависимости

1. формирование разбиений вещественной оси на интервалы Δ_k, $k = 1, \ldots, n$;
2. задание разбиений $\mathbf{K}^i$, $i = 1, \ldots, N$, формирование обучающих выборок задач распознавания Z_i, $i = 1, \ldots, N$, выбор классификаторов A_i, $i = 1, \ldots, N$, их обучение;
3. вычисление вектора весов $\vec{w}$.

Алгоритм вычисления значения зависимой величины

1. классификация алгоритмами $A_1, \ldots, A_N$ объекта $\vec{x}$;
2. вычисление вектора ответов $\vec{u}(\vec{x})$;
3. вычисление значения $f(\vec{x})$.

Теорема 1.3 *Линейный корректор при $n = m'$ является корректной моделью.*

Утверждение 1.5 *Для случая $L = 2, N = n - 1$ и $n = m'$ вектор весов имеет вид: $w_k = \frac{y_k - y_{k-1}}{u_k^{(2)} - u_k^{(1)}}$, $k = 2, \ldots, n$, $w_1 = y_1 - \sum\limits_{k=2}^{n} u_k^{(1)} w_k$.*

Для обучения корректной модели восстановления зависимости с разбиением области значений целевого признака на m' интервалов необходимо обучить $m' - 1$ классификаторов, т.е. число классификаторов сравнимо с количеством объектов обучающей выборки.

Погрешность модели восстановления зависимости есть $\sum\limits_{i=1}^{m} |y_i - c_{k_i}|^2$, $y_i \in \Delta_{k_i}$, $i = 1, \ldots, m$. Данное выражение совпадает с погрешностью разбиения $Q(\mathbf{\Delta})$, т.е. для построения модели с заданной погрешностью δ необходимо построить разбиение с погрешностью δ.

Теорема 1.4 *Линейный корректор при $n < m'$ является квазикорректной относительно разбиения $\mathbf{\Delta}$ моделью.*

Во второй главе рассматривается свойство устойчивости моделей «байесовский корректор» и «линейный корректор». Устойчивость является аналогом регулярности по Журавлеву для распознающих алгоритмов.

Под моделью A будем понимать байесовский или линейный корректор с корректными классификаторами.

Введем обозначение $A(Z_m) = \vec{\hat{y}} = (\hat{y}_1, \ldots, \hat{y}_m)^T$.

Рассмотрим пространство задач с метрикой $\rho(Z_m, Z'_m)$. Обозначим $\mathcal{Z}_\delta(Z_m) = \{Z'_m | \rho_X(Z_m, Z'_m) < \delta\}$ — δ-окрестность задачи Z. Введем метрику на пространстве алгоритмов $\rho_A(A(Z_m), A(Z'_m)) = \|\vec{\hat{y}} - \vec{\hat{y}}'\|$.

Определение 12 *Модель восстановления зависимости A устойчива, если* $\forall \varepsilon > 0, \exists \delta = \delta(\varepsilon) : \forall Z'_m \in \mathcal{Z}_\delta(Z_m) \Rightarrow \rho_A(A(Z_m), A(Z'_m)) < \varepsilon$.

В разделе 2.1 рассматривается свойство устойчивости моделей относительно изменения описаний объектов.

Рассмотрим задачу с изменением описаний объектов $Z'_m = \{(\vec{x}_i{}', y_i)\}_{i=1}^m$. Разбиение области значений целевого признака $\mathbf{\Delta}'$ задачи Z'_m для модели A совпадает с разбиением $\mathbf{\Delta}$, т.к. значения целевого признака $\vec{y}$ одинаковы для задач Z_m и Z'_m. При этом оценки значений целевого признака для задач Z_m и Z'_m совпадают, $\vec{\hat{y}} = \vec{\hat{y}}'$.

Рассмотрим пространство задач с метрикой $\rho_X(Z_m, Z'_m) = \sqrt{\sum_{i=1}^m \|\vec{x}_i - \vec{x}_i{}'\|^2}$.

Обозначим $m_y(Z)$ - число различных значений целевого признака задачи Z. Обозначим $A^*(Z)$ - модель для задачи Z с разбиением $\mathbf{\Delta} : \|\mathbf{\Delta}\| = m_y(Z)$. Напомним, что модель $A^*(Z_m)$ корректна, $\vec{\hat{y}} = \vec{y}$. Также корректна модель $A^*(Z'_m)$, $\vec{\hat{y}}' = \vec{y}'$.

Теорема 2.5 *Модель A^* устойчива относительно изменения описаний объектов.*

В разделе 2.2 рассматривается свойство устойчивости моделей относительно изменения значений целевого признака объектов.

Рассмотрим задачу с изменением значений целевого признака объектов $Z'_m = \{(\vec{x}_i, y'_i)\}_{i=1}^m$. Рассмотрим пространство задач с метрикой $\rho_Y(Z_m, Z'_m) = \|\vec{y} - \vec{y}'\|$. δ-окрестность задачи Z_m есть $\mathcal{Z}_\delta(Z_m) = \{Z'_m | \rho_Y(Z_m, Z'_m) < \delta\}$.

Теорема 2.6 *Модель A^* устойчива относительно изменения значений целевого признака объектов.*

В разделе 2.3 рассматривается свойство устойчивости моделей относительно изменения описаний и значений целевого признака объектов.

Рассмотрим задачу с изменением описаний и значений целевого признака объектов $Z'_m = \{(\vec{x}_i{}', y'_i)\}_{i=1}^m$. Рассмотрим пространство задач с метрикой $\rho_{XY}(Z_m, Z'_m) = \sqrt{\sum_{i=1}^m \left(w_x \|\vec{x}_i - \vec{x}_i{}'\|^2 + w_y(y_i - y'_i)^2\right)}$, где $w_x, w_y > 0$.

Теорема 2.7 *Модель A^* устойчива относительно изменения описаний и значений целевого признака объектов.*

Модель A с разбиением $\mathbf{\Delta} : \|\mathbf{\Delta}\| < m_y(Z)$ не обладает свойством устойчивости.

В третьей главе приводятся результаты практической апробации предложенных моделей. Производится сравнение результатов моделей «байесовский корректор» и «линейный корректор» и известных моделей восстановления зависимостей на реальных задачах.

В качестве классификаторов в моделях «байесовский корректор» и «линейный корректор» использовался алгоритм голосования по системам логических закономерностей классов.

Описание задач:

1. «медицинская задача» — описание объекта состоит из 30 бинарных и количественных признаков: медицинские показатели состояния пациента, зависимая величина — частота кризов артериальной гипертензии; множество объектов разбито на непересекающиеся обучающую (263 объекта) и контрольную (66 объектов) выборки;

2. «стоимость домов» — описание объекта состоит из 13 бинарных и количественных признаков: характеристики дома и района, в котором он расположен, зависимая величина — стоимость дома; множество объектов разбито на непересекающиеся обучающую (253 объекта) и контрольную (253 объекта) выборки;

3. «прочность бетона» — описание объекта состоит из 8 количественных признаков: массовые доли компонент и возраст бетона, зависимая величина — прочность бетона; множество объектов разбито на непересекающиеся обучающую (772 объекта) и контрольную (257 объектов) выборки.

Для сравнения моделей использовалась среднеквадратичная ошибка $MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ для нормированных значений целевого признака.

Задача	Байес.корр.	Лин.корр.	Линейная регрессия	Ядерное сглаживание
Мед. задача	0.00417	0.0021	0.02011	0.00701
Стоимость домов	0.00229	0.00111	0.01034	0.00679
Прочность бетона	0.01551	0.01127	0.02117	0.01745

Модели «байесовский корректор» и «линейный корректор» дали меньшую среднеквадратичную ошибку на всех рассмотренных задачах.

В заключении сформулированы основные результаты диссертационной работы.

Основные результаты:

1. Разработан общий подход к формированию задач распознавания и вычисления значения зависимой величины как коллективного решения.
2. Разработаны модели восстановления зависимостей «байесовский корректор» и «линейный корректор».
3. Доказаны свойства корректности и устойчивости моделей восстановления зависимостей «байесовский корректор» и «линейный корректор».
4. Подтверждена применимость разработанных моделей восстановления зависимостей для решения реальных прикладных задачах результатами практической апробации.

Список публикаций

[1] Рязанов В. В., Ткачев Ю. И. Решение задачи восстановления зависимости коллективами распознающих алгоритмов // Доклады 14-й Всероссийской

конференции «Математические методы распознавания образов» ММРО-2009. — М.: МАКС Пресс, 2009. — С. 172-175.

[2] Рязанов В. В., Ткачев Ю. И. Решение задачи восстановления зависимости коллективами распознающих алгоритмов // Доклады Академии наук. 2010. Т. 432. № 6. С. 750-754.

[3] Рязанов В. В., Ткачев Ю. И. Восстановление зависимости на основе байесовской коррекции коллектива распознающих алгоритмов // ЖВМиМФ. 2010. Т. 50, № 9. С. 1687-1696.

[4] V.V. Ryazanov, Ju.I. Tkachev. Restoring of Dependences of Samples of Precedents with Usage of Models of Recognition // New Trends in Classification and Data Mining, ISBN 978-954-16-0042-9, ITHEA, Sofia, Bulgaria, 2010, pp.17-24.