

Рецензенты: член-корреспондент РАН, доктор технических наук, профессор М. М. Мирошников и кафедра компьютерной фотоники Санкт-Петербургского государственного университета информационных технологий, механики и оптики

П64 Потанов А. С.
Распознавание образов и машинное восприятие:

Общий подход на основе принципа минимальной длины описания. — СПб.: Политехника, 2007. — 548 с.: ил.

ISBN 5-7325-0881-3

В книге подробно рассмотрены принципы минимальной длины описания, являющийся следствием теоретико-информационного подхода к построению моделей и выбору гипотез. Этот принцип становится все более популярным при решении сложных задач автоматического анализа данных, традиционно относившихся к области искусственного интеллекта. Рассмотрены задачи распознавания образов, машинного восприятия и грамматического вывода, для которых используются принципиально минимальной длины описания уже позволяло получить более эффективные решения. На конкретных примерах показана возможность разработки унифицированного подхода к решению указанных задач.

Книга предназначена для широкого круга читателей: студентов, молодых ученых и специалистов, интересующихся компьютерными науками и, в частности, искусственным интеллектом.

ISBN 5-7325-0881-3

© А. С. Потанов, 2007
© Издательство «Политехника», 2007

УДК 004.855
ББК 32.973.26

ОГЛАВЛЕНИЕ

| | |
|---|-----------|
| Предисловие | 8 |
| Глава 1 | |
| ИНДУКТИВНЫЙ ВЫВОД | 11 |
| 1.1. Проблема выбора типотез в индуктивном выводе | |
| 1.1.1. Что такое индуктивный вывод? Неформальное рассмотрение | 11 |
| 1.1.2. Основные понятия индуктивного вывода | 13 |
| 1.1.3. Критерии сравнения гипотез | 15 |
| 1.1.4. Бритва Оккама и принцип минимальной длины описания | 18 |
| 1.1.5. Бритва Оккама в научной эстетике и биологических системах | 20 |
| 1.2. Байесовские методы в индуктивном выводе и машинном обучении | 23 |
| 1.2.1. Теорема Байеса для выбора модели | |
| 1.2.2. Принятие решений и предсказание на основе правды Байеса | 27 |
| 1.2.3. Методы максимума апостериорной вероятности и максимального правдоподобия | 29 |
| 1.2.4. Проблема априорных вероятностей | 31 |
| 1.3. Основные положения теории информации | 39 |
| 1.3.1. Теория информации Шеннона: историческая справка | |
| 1.3.2. Энтропия дискретной случайной величины | 41 |
| 1.3.3. Энтропия непрерывной случайной величины | 45 |
| 1.3.4. Префиксное кодирование | 49 |
| 1.4. Информационная мера при выборе модели | 57 |
| 1.4.1. Теоретико-информационная интерпретация правила Байеса | |
| 1.4.2. Методы второго порядка и предположение нормальности | 61 |
| 1.4.3. Среднеквадратичное отклонение и энтропия | 64 |
| 1.4.4. Коэффициент корреляции и средняя взаимная информация | 69 |
| 1.4.5. Проблема информативности модели | 74 |
| 1.5. Машинная Тьюринга и алгоритмическая сложность | 76 |
| 1.5.1. Понятие алгоритма | |
| 1.5.2. Формализм машины Тьюринга | 78 |
| 1.5.3. Универсальная машина Тьюринга | 80 |
| 1.5.4. Понятие алгоритмической сложности | 83 |

| | |
|---|-----|
| 1.5.5. Индивидуальная случайность бинарной строки | 85 |
| 1.5.6. Алгоритмическая сложность как количество информации | 89 |
| 1.6. Алгоритмическая сложность и сравнение гипотез | 91 |
| 1.6.1. Предказание на основе алгоритмической вероятности | — |
| 1.6.2. Алгоритмическая сложность в индуктивном выводе | 94 |
| 1.6.3. Индукция и предсказание | 96 |
| 1.6.4. Полнота и комбинаторный взрыв | 98 |
| 1.6.5. Проблема субъективности и инкрементное машинное обучение | 102 |
| 1.7. Заключение | 106 |
| Глава 2 | |
| НИЗКОУРОВНЕВЫЕ ЗАДАЧИ МАШИННОГО ОБУЧЕНИЯ | |
| 2.1. Распознавание образов в контексте машинного обучения | — |
| 2.1.1. Вводные замечания по проблеме машинного обучения | — |
| 2.1.2. Основные понятия распознавания образов | 115 |
| 2.1.3. Дополнительные предположения о пространстве описаний и множестве классов | 118 |
| 2.1.4. Постановка задачи распознавания в зависимости от количества априорной информации | 122 |
| 2.1.5. Задачи распознавания в терминах индуктивного вывода | 128 |
| 2.2. Классификация образов | 130 |
| 2.2.1. Решающие функции | — |
| 2.2.2. Критерии, основанные на функциях расстояния | 135 |
| 2.2.3. Статистический подход | 138 |
| 2.2.4. Информационный критерий | 141 |
| 2.3. Распознавание с учителем | 144 |
| 2.3.1. Линейные решающие функции и опорные векторы | — |
| 2.3.2. Обобщенные решающие функции и ядра | 148 |
| 2.3.3. Выбор эталонных образов | 152 |
| 2.3.4. Параметрические методы оценивания плотности вероятности | 155 |
| 2.3.5. Параметрические методы оценивания плотности вероятности | 160 |
| 2.3.6. Информационные критерии в распознавании | 163 |
| 2.3.7. Принцип МДО и априорные ограничения методов распознавания | 171 |
| 2.3.8. Пример практического приложения: распознавание целей | 176 |

| | |
|---|-----|
| 2.4. Группирование образов в пространстве признаков | 180 |
| 2.4.1. Проблема обучения без учителя | — |
| 2.4.2. Задача группирования | 183 |
| 2.4.3. Кластеризация на основе функций расстояния | 185 |
| 2.4.4. Использование смесей в задаче группирования | 191 |
| 2.4.5. Критерии выбора числа кластеров | 197 |
| 2.4.6. Основные упрощения в постановке задачи группирования | 202 |
| 2.5. Выбор признаков | 205 |
| 2.5.1. Общие замечания о проблеме выбора признаков | — |
| 2.5.2. Преобразование кластеризации при обучении с учителем | 208 |
| 2.5.3. Проблема выбора признаков при обучении без учителя | 214 |
| 2.5.4. Анализ главных компонент и факторный анализ | 216 |
| 2.5.5. Уменьшение избыточности данных и поиск интересных направлений в пространстве признаков | 222 |
| 2.5.6. Анализ независимых компонент | 224 |
| 2.5.7. Представления информации, объединяющие свойства распределенных и локальных представлений | 228 |
| 2.5.8. Информационный критерий качества представления | 229 |
| 2.5.9. Пример практического приложения: выбор текстовых признаков | 234 |
| 2.6. Регрессия и сегментация | 241 |
| 2.6.1. Задача регрессии | — |
| 2.6.2. Проблема выбора факторов и ее решение с помощью принципа МДО | 243 |
| 2.6.3. Задача сегментации | 247 |
| 2.6.4. Информационный критерий качества сегментации | 249 |
| 2.7. Заключение | 252 |
| Глава 3 | |
| МАШИННОЕ ВОСПРИЯТИЕ | |
| 3.1. Представление изображений в системах компьютерного зрения | — |
| 3.1.1. Машинное восприятие в контексте искусственного интеллекта | — |
| 3.1.2. Интерпретация изображений как центральная проблема компьютерного зрения | 260 |
| 3.1.3. Представления в виде необработанных данных: пиксельный уровень | 263 |
| 3.1.4. Низкоуровневые представления: математические модели изображений | 265 |
| 3.1.5. Средний уровень: структурные методы | 272 |

| | |
|--|-----|
| 3.1.6. Верхний уровень: методы, основанные на знаниях | 281 |
| 3.1.7. Иерархические представления изображений | 285 |
| 3.2. Принципы минимальной длины описания в интерпретации изображений | 291 |
| 3.2.1. Выбор представления изображений с теоретико-информационной точки зрения | — |
| 3.2.2. Общие предположения о свойствах изображений | 295 |
| 3.2.3. Семантика изображений на одномерной области | 300 |
| 3.2.4. Построение структурных элементов на основе контурной информации | 310 |
| 3.2.5. Формирование составных структурных элементов | 316 |
| 3.2.6. Пример практического приложения: совмещение изображений | 328 |
| 3.2.7. Некоторые выгоды относительно общей проблемы индукции | 334 |
| 3.3. Теоретико-информационный подход к машинному восприятию речи | 335 |
| 3.3.1. Проблема машинного слуха и распознавание речи | — |
| 3.3.2. Основные понятия в области распознавания речи | 338 |
| 3.3.3. Распознавание фонем по различительным признакам | 340 |
| 3.3.4. Распознавание слов по цепочкам символов | 347 |
| 3.3.5. Выделение границ слов и модели языка на основе N-грамм | 352 |
| 3.3.6. Выделение устойчивых сочетаний фонем | 357 |
| 3.3.7. Ограничения рассмотренных методов машинного восприятия | 366 |
| 3.4. Формирование лингвистических единиц, основанных на семантике, на примере системы CELL | 368 |
| 3.4.1. Проблема смысла референтных выражений | — |
| 3.4.2. Общая архитектура системы CELL | 371 |
| 3.4.3. Реализация зрительной и акустической подсистем в системе CELL | 376 |
| 3.4.4. Основные результаты тестирования системы CELL | 378 |
| 3.4.5. Дальнейшее развитие системы CELL | 380 |
| 3.4.6. Нерешенные проблемы автоматического построения концептуальных систем | 381 |
| 3.5. Иерархические представления, неподлая декомпозиция задач и адаптивный резонанс | 390 |
| 3.5.1. Введение иерархичности при решении NP-полных задач | — |
| 3.5.2. Понятие адаптивного резонанса | 392 |
| 3.5.3. Теоретико-информационная интерпретация адаптивного резонанса | 394 |
| 3.5.4. Адаптивный резонанс при интерпретации изображений | 396 |
| 3.5.5. Адаптивный резонанс в анализе речи | 400 |

| | |
|---|------------|
| интерпретации аудио- и видео информации | 403 |
| 3.5.7. Концепция метасистемных переходов | 407 |
| 3.6. Заключение | 410 |
| Глава 4 | |
| ВЫСОКОУРОВНЕВЫЕ ЗАДАЧИ ИНДУКТИВНОГО ВЫВОДА | 413 |
| 4.1. Проблема индуктивного вывода символьных представлений | — |
| 4.2. Формальные грамматики | 419 |
| 4.2.1. Историческая справка | — |
| 4.2.2. Основные определения | 421 |
| 4.2.3. Типы формальных грамматик | 428 |
| 4.2.4. Стохастические грамматики | 431 |
| 4.2.5. Синтаксический разбор | 434 |
| 4.3. Грамматический вывод | 439 |
| 4.3.1. Основные определения и постановка задачи | — |
| 4.3.2. Восстановление грамматики перечислением | 443 |
| 4.3.3. Эвристические процедуры грамматического вывода | 446 |
| 4.3.4. Байесовский вывод стохастических грамматик | 452 |
| 4.3.5. Теоретико-информационный подход к грамматическому выводу | 454 |
| 4.3.6. Некоторые замечания о восстановлении грамматики при информаторном представлении | 463 |
| 4.4. Приложения методов восстановления грамматик на основе принципа МДО в анализе естественных языков | 467 |
| 4.4.1. Краткое сравнение формальных грамматик с моделями языка на основе N-грамм | — |
| 4.4.2. Обучение фразам | 470 |
| 4.4.3. Разделение морфов на классы на основе принципа МДО | 474 |
| 4.4.4. Построение классов слов на основе принципа МДО | 478 |
| 4.4.5. Проблема выделения подзадач при восстановлении грамматики | 483 |
| 4.5. Наборы правил, деревья и графы решений | 488 |
| 4.5.1. Построение наборов порождающих правил | — |
| 4.5.2. Информационный критерий качества дерева решений | 498 |
| 4.5.3. «Жадные» алгоритмы построения деревьев решений | 506 |
| 4.5.4. Ограничения представления информации в форме деревьев решений | 514 |
| 4.5.5. Представления, расширяющие деревья решений | 517 |
| 4.5.6. Обусловленные символьных представлений | 522 |
| 4.6. Заключение | 525 |
| Литература | 527 |

Задача поиска закономерностей в некотором наборе исходных данных возникает на всех уровнях работы мозга, начиная с сенсорного восприятия и заканчивая расцучной деятельностью. В науке также существует проблема анализа эмпирических данных, которая является одной из центральных и обычно обозначается как проблема индуктивного вывода. Индуктивный вывод часто формулируется как проблема выбора модели, наилучшим образом описывающей или объясняющей имеющиеся данные. И здесь возникает вопрос: что же является критерием качества модели? Именно этот вопрос является центральным для всей книги.

При автоматизации различных сфер человеческой деятельности проблема индуктивного вывода оказывается одной из главных: построение машинной системы, которая бы проявляла свойства адаптивности и обучаемости, невозможно без ее решения. Осуществление прогнозирования также невозможно без анализа эмпирических данных.

В зависимости от природы исходных данных задача их анализа может превращаться в задачу анализа изображений, акустического сигнала или сигналов сенсоров других типов, лингвистического анализа, распознавания образов или машинного обучения вообще. Для каждой из соответствующих (а также многих других) предметных областей разработано множество методов автоматического анализа. Как правило, эти методы опираются на частные эвристики, работающие лишь при определенных ограничениях на исходные данные. Однако на практике эти ограничения могут не соблюдаться, что приводит к ухудшению результатов анализа. Наиболее явно эта проблема проявляется в использовании эвристических критериев качества модели. Такие критерии, хотя и выглядят вполне корректными, могут приводить к не вполне адекватному выбору модели. Чаще всего эта неадекватность проявляется в проблеме переобучения, или чрезмерно близкой подгонки, о которой будет подробно говориться в книге. В связи с этим для каждой предметной области существует необходимость разработки таких методов анализа, которые бы имели теоретическую основу и, в частности, использовали стро-го обоснованный критерий качества модели.

Хотя исходные данные могут очень сильно различаться в разных предметных областях, сама задача анализа дан-

ных сохраняется практически без изменений. Достаточно часто бывает так, что методы, разработанные для одного типа данных, оказывались применимыми для другого типа. Создание общей методики, которая бы формально описывала процесс разработки алгоритмов автоматического анализа данных независимо от их типа, является, безусловно, весьма актуальным.

В этой книге рассматривается о теоретико-информационном подходе к проблеме индуктивного вывода. Основные его идеи были сформулированы в 1960-х годах. Центральная концепция этого подхода выражается в принципе минимальной длины описания (МДО): среди всех моделей следует выбрать ту, которая позволяет описать данные наиболее коротко (с учетом длины описания самой модели), причем длины описаний определялись через алгоритмическую сложность. Этот подход давал строгое теоретическое разрешение проблемы критерия качества модели в индуктивном выводе, однако исходно не мог быть применен на практике из-за ряда трудностей. Предпосылки возникновения данного подхода, его теоретическая основа, а также трудности, связанные с практическим применением, описаны в гл. 1.

Принцип МДО получил широкое распространение при решении задач автоматического анализа данных только в 1990-х годах. В различных областях информатики этот принцип неоднократно переизобретался в более частных и конкретных формах, однако с сохранением общей идеи. Обширная литература, посвященная использованию принципа МДО, показывает его явные преимущества в получении более эффективных решений по сравнению с другими методами. При этом принцип МДО не отвергает существующие методы, а позволяет уточнить использующиеся в них критерии качества модели, что приводит к улучшению этих методов. В то же время далеко не всегда является очевидным, как именно следует применять этот принцип.

Автором предпринимается попытка обобщить существующий мировой опыт по применению принципа МДО в задачах анализа данных разных типов. Проблемам распознавания образов, регрессии и сегментации и их решению на основе принципа МДО посвящена гл. 2 книги. В гл. 3 описывается применение этого принципа для решения задач анализа изображений и речи. Проблемы теоретико-информационного подхода в задачах восстановления траматик и деревьев решений рассматриваются в гл. 4. В книге при-

1.1. ПРОБЛЕМА ВЫБОРА ГИПОТЕЗ В ИНДУКТИВНОМ ВЫВОДЕ

1.1.1. Что такое индуктивный вывод? Неформальное рассмотрение

Все рациональные рассуждения традиционно делятся на дедуктивные и индуктивные [1, с. 141]. Принято считать, что индукция — это умозаключение от частных фактов к некоему общему гипотетическому утверждению, в то время как дедукция — это способ рассуждения, при котором осуществляется переход от общего знания или фактов к частным следствиям. Однако индуктивному выводу придается и более широкий смысл, если рассмотрение не ограничивается формальной логикой. Наиболее широко индуктивный вывод можно определить как проблему выбора модели из некоторого множества моделей, которая наилучшим образом «объясняет» исходные данные [2, с. 1]. Здесь под частными фактами понимается набор данных, а под общим утверждением — модель, описывающая эти данные (содержащаяся в них закономерности).

Это означает, что индуктивным выводом будет являться и проведение некоторой интерполяционной кривой по заданному набору точек, и составление словесного описания изображения. Более того, многие виды реальных рассуждений, традиционно относимых к дедуктивным, могут также быть причислены и к индуктивным. Так, классический пример дедуктивного рассуждения [1, с. 143]: «Все люди смертны. Сократ — человек, следовательно, Сократ смертен» описывается на две послылки, по крайней мере, первая из которых («Все люди смертны») не является с необходимостью истинной, а является результатом обобщения данных опыта. Тогда и консеквенту (заключению) этого вывода («Сократ смертен») можно приписать лишь некоторую вероятность, отличную от единицы, а значит, в этом выводе производится выбор более достоверной гипотезы из двух возможных. Именно недостоверность результата и является признаком индуктивного вывода, отличающим его от де-

водятся некоторые рекомендации по построению автоматических методов анализа, которые являются обобщением опыта, накопленного исследователями в данной области.

Автор надеется, что этот материал даст читателю достаточно полное представление о сути теоретико-информационного подхода и поможет успешно применить его на практике при решении задач из своей предметной области. Таким образом, книга может оказаться полезной для разработчиков прикладных систем автоматического анализа данных. В книге также затрагиваются теоретические вопросы индуктивного вывода, которые могут оказаться интересными научным работникам, специализирующимся на некоторых разделах искусственного интеллекта. Обсуждение возможности применения принципа МДО в распознавании образов, анализе изображений и речи, лингвистическом анализе и т. д. сопровождается достаточно подробным описанием классических методов, используемых в этих областях.

Автор выражает благодарность С. А. Родионову, Н. И. Потаповой, В. А. Дяховцевскому и И. Г. Гарипову за то, что они взяли на себя труд ознакомиться с рукописью и высказали ряд конструктивных замечаний. Автор признателен А. Плахову, с которым планировалось совместное написание книги, чему, к сожалению, помешали жизненные обстоятельства. Автор также благодарен коллективу лаборатории автоматических методов обработки изображений Государственного Оптического Института им. С. И. Вавилова, опыт работы в котором сыграл существенную роль при написании данной книги.

Все замечания и пожелания автор просит направлять по адресу: 191023, Санкт-Петербург, Инженерная ул., 6, издательство «Политехника» или на электронный адрес автора: protarov@mail.wrlus.net