АКАДЕМИЯ НАУК СССР
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

Распознавание
Классификация
Прогноз

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И ИХ ПРИМЕНЕНИЕ

Выпуск 1

Ежегодник основан в 1988 г.

Ответственный редактор
член-корреспондент АН СССР
Ю. И. ЖУРАВЛЕВ

МОСКВА «НАУКА» 1989

ВВЕДЕНИЕ

ОБ АЛГЕБРАИЧЕСКИХ МЕТОДАХ В ЗАДАЧАХ РАСПОЗНАВАНИЯ И КЛАССИФИКАЦИИ

Ю. И. ЖУРАВЛЕВ

Проблема распознавания в течение достаточно продолжительного времени привлекает внимание специалистов в области прикладной математики, а затем и информатики. Отметим, в частности, работы Р. Фишера, выполненные в 1920-х годах и приведшие к формированию дискриминантного анализа, постановку вначале 1940-х годов А. Н. Колмогоровым и А. Я. Хинчиным задачи о разделении смеси двух распределений, теорию статистических решений и массу работ 1950—1960-х годов, посвященных поиску и применению алгоритмов, обеспечивающих отнесение нового объекта к одному из заданных классов или разделение некоторого множества объектов на несколько непересекающихся классов. К середине 1970-х годов облик распознавания как самостоятельного научного направления стал несколько меняться, поскольку оно достигло такой стадии развития, что возникла возможность создания нормальной математической теории распознавания.

Одной из предпосылок этой возможности явилось выделение и отработка в процессе решения прикладных задач обработки информации ряда моделей алгоритмов распознавания — семейств алгоритмов для решения классификационных задач. К этому времени были изучены и получили практическое распространение главным образом следующие модели.

1. Модели, основанные на использовании принципа разделения (R-модели). Эти модели различаются главным образом заданием класса поверхностей, среди которых выбирается поверхность (или набор поверхностей), в некотором смысле наилучшим образом разделяющая элементы разных классов.

2. Статистические модели. Этот тип моделей алгоритмов распознавания основан на использовании аппарата математической статистики. Они применяются в основном в тех случаях, когда известны или могут быть просто определены вероятностные характеристики классов, например соответствующие функции распределения.

3. Модели, построенные на основе так называемого метода потенциальных функций (П-модели). В основе этой модели лежит заимствованная из физики идея потенциала, определенного для любой точки пространства и зависящего от того, где расположен

источник потенциала. В качестве функции принадлежности объекта классу используется потенциальная функция — всюду положительная и монотонно убывающая функция расстояния.

4. Модели вычисления оценок (голосования) (Г-модели). Эти модели основаны на принципе частичной прецедентности. Анализируется «близость» между частями описаний ранее классифицированных объектов и объекта, который надо распознать. Наличие близости служит частичным прецедентом и оценивается по некоторому заданному правилу (посредством числовой оценки). По набору оценок близости вырабатывается общая оценка распознаваемого объекта для класса, которая и является значением функции принадлежности объекта классу.

5. Модели, основанные на исчислении высказываний, в частности на аппарате алгебры логики (Л-модели). В этих моделях классы и признаки объектов рассматриваются как логические переменные, а описание классов на языке признаков представляется в форме булевых соотношений.

Было бы естественно считать, что центральной задачей теории распознавания образов является разработка эффективных вычислительных средств для отнесения формализованных описаний объектов распознавания к соответствующим классам. В основе такого отнесения (распознавания, классификации) лежит получение некоторой агрегированной оценки объекта, исходя из его описания. Столь же естественно считать, что задачи распознавания — это дискретные аналоги задач поиска оптимальных решений. В этом случае можно говорить о широком классе задач, в которых по некоторой, обычно весьма разнородной, быть может, неполной, нечеткой. противоречивой, искаженной и косвенной информации требуется установить. обладают ли изучаемые (весьма сложные, в некотором смысле «комплексные») ситуации (объекты, явления) фиксированным конечным набором свойств, позволяющим отнести их к определенному классу — задачи распознавания и классификации, или по аналогичного рода информации о конечном множестве достаточно однотипных процессов следует выяснить, в какой из конечного числа областей будут находиться эти процессы через определенный период времени—задачи прогноза. Практическое значение постановки и решения задач такого рода и очень нетрудно привести множество примеров содержательных задач, сводящихся к распознавательской постановке, из человеческой деятельности практически любого рода.

Внешне достоинства, достижения и перспективы распознавания воспринимаются именно в такой перспективе. Не менее существенным является, однако, и другая сторона дела. Она заключается в том, что становление распознавания служит отличной моделью развития математической теории обработки и преобразования информации, развития, в процессе которого эвристические (по крайней мере, по существу) методы получают строгое обоснование и начинают применяться в рамках вполне формализованных регулярных процедур. Интересно отметить, что само распознава-

ние сегодня является достаточно разработанным вариантом такой теории, поскольку позволяет разрешать ее основную задачу — синтезировать и выбирать алгоритмические средства для извлечения полезной информации из данных того рода, который был охарактеризован выше.

Известно, что к постановке задачи распознавания прибегают в тех случаях, когда трудно строить формальные теории и применять классические математические методы, что происходит обычно в силу одной из следующих двух причин: а) уровень формализации соответствующей предметной области и/или доступная информация таковы, что не могут составить основу для синтеза математической модели, отвечающей классическим математическим или математико-физическим канонам и допускающей изучение классическими аналитическими или численными методами; б) математическая модель, в принципе, может быть построена, однако ее синтез или изучение связаны с такими затратами, что они существенно превышают выигрыш, приносимый искомым решением, либо выходят за пределы существующих технических возможностей, либо делают решение задачи просто бессмысленным.

Таким образом, «двойственность» распознавания проявилась в том, что решение таких задач ввело в обиход большое число некорректных (эвристических) алгоритмов. Довольно долго подавляющее большинство приложений теории распознавания было связано с плохо формализованными областями — медициной, геологией, социологией, химией и т. д. Сегодня в этих областях еще трудно строить формальные теории и применять стандартные математические методы. В лучшем случае удается дать математическое оформление некоторым интуитивным принципам и затем применять полученные «эмпирические формализмы» для решения частных задач. Это обстоятельство определило тот факт, что на первом этапе развития теории и практики распознавания возникло большое число различных методов и алгоритмов, применявшихся без какого-либо серьезного обоснования для решения практических задач. При исследовании задачи или класса задач на базе так называемых «правдоподобных» рассуждений предлагался нестрогий, но содержательно разумный метод решения и основанный на нем алгоритм; обоснование же производилось непосредственно в эксперименте с задачами. Алгоритмы, выдержавшие подобную экспериментальную проверку, т. е. приносившие успех при решении определенных практических задач, применяются, несмотря на отсутствие математических обоснований.

Стало очевидным, что появление каждого эвристического алгоритма такого рода можно рассматривать как некоторый эксперимент, а со всем множеством экспериментов и их результатов — работать как с новым для математики множеством объектов, т. е. изучать с помощью строгих математических методов множество некорректных процедур решения плохо формализованных задач. Поэтому второй этап развития теории распознавания отличался, с одной стороны, попытками ставить и решать задачу выбора в кон-

кретной ситуации наилучшего в некотором смысле алгоритма, с другой — попытками переходить от описания отдельных некорректных алгоритмов к описанию принципов их формирования, т. е. попытками строить единообразные описания для множеств эвристических, но успешно решающих реальные задачи процедур. Подобное множество задается указанием переменных, объектов, функций, параметров и точным определением областей их вариации. Фиксация этих переменных, объектов, функций, параметров позволяет выделить из соответствующего множества, т. е. модели, некоторый конкретный алгоритм. Впервые в виде модели был представлен класс алгоритмов вычисления оценок, позднее появились описания и других моделей.

Потребность в синтезе моделей алгоритмов распознавания в первую очередь определялась необходимостью фиксировать каким-то образом класс алгоритмов при выборе оптимальной или хотя бы приемлемой процедуры решения конкретной задачи. В свою очередь, попытки построения таких моделей породили интерес к собственно «математическим» свойствам алгоритмов распознавания и в особенности к проблемам их строгого обоснования. Оказалось, что получение описания класса алгоритмов распознавания представляет собой задачу, сходную с построением классического определения алгоритма. Следовательно, необходимым условием построения теории распознавания являлось проведение классических алгоритмических исследований для понятия «алгоритм распознавания».

Анализ совокупности некорректных алгоритмов распознавания позволяет по мере их накопления выделять и описывать не только отдельные частные алгоритмы, но и принципы их формирования. Эти принципы, действующие уже над подмножествами алгоритмов и формулируемые сначала также в плохо формализованном виде, затем могут реализовываться в виде точных математических описаний. На этом этапе эвристический характер имеет собственно выбор принципа, а алгоритмы, порождаемые на основе соответствующего принципа, могут строиться стандартным образом. Именно в таком смысле формализация различных принципов построения распознающих алгоритмов приводит к появлению моделей распознающих алгоритмов.

Переход к моделям распознающих алгоритмов сам по себе не привел ни к созданию некоей универсальной модели, ни к формализации процесса выбора определенной модели для решения конкретной задачи распознавания. Но появление моделей позволило ставить и решать в рамках определенной модели задачу выбора алгоритма, экстремального по функционалу качества классификации или прогноза. Построение таких оптимальных алгоритмов обычно приводит к исследованию, реализации и разработке вычислительных схем для нестандартных экстремальных задач.

Параметризация ряда алгоритмов (моделей) распознавания и возможность на основе имеющейся информации о классах определять значения параметров действительно позволяет выбирать

корректные алгоритмы для некоторых подклассов задач. В большинстве практических случаев, однако, оказывается, что подкласс этот довольно узок, так как в противном случае при синтезе модели алгоритмов распознавания, описании классов и выборе признаков объектов распознавания необходимо было бы использовать весьма значительный объем априорной информации, которую можно получить, лишь располагая достаточно точной моделью изучаемых объектов и явлений. Кроме того, построение оптимального алгоритма в многопараметрической модели связано с решением трудных экстремальных задач (часто NP-полных). Достаточно часто не удается отыскать глобальный экстремум, использование же алгоритмов, соответствующих локальному экстремуму, значительно ухудшает качество распознавания и не позволяет реализовать истинные возможности модели. Иногда оказывается, что использование малопараметрических моделей, допускающих отыскание глобального экстремума, дает больший эффект, чем применение локально-экстремального алгоритма из многопараметрической модели; нет к тому же и гарантии, что оптимальный в модели алгоритм останется таким же и при работе с объектами, не участвовавшими в обучении.

Обоснование на втором этапе проводится одним из следующих трех способов:

1) экспериментально — возможность получить некое приемлемое с точки зрения пользователя «решение» поставленной задачи с помощью соответствующего алгоритма распознавания рассматривается в качестве обоснования допустимости его использования при решении данной задачи;

2) при помощи решения оптимизационной задачи и использования оптимального в рамках выбранной модели алгоритма распознавания — обоснование состоит в том, что применяется наилучший из возможных в используемой модели алгоритмов распознавания;

3) обоснование проводится так же, как и в п. 2, но к тому же доказывается, что при выполнении ряда «естественных» гипотез (условий), справедливых для изучаемого класса задач, алгоритмы, оптимальные в используемой модели, действительно обеспечивают высокую точность распознавания, т. е. обосновывается как выбор алгоритма, так и выбор модели.

Очередной этап развития распознавания был связан с изучением строения совокупности некорректных алгоритмов в целом. Поскольку оказалось, что обогащение модели часто не удается сопроводить эквивалентным улучшением результатов и к тому же существует естественная граница сложности любой модели, возникла идея выбирать алгоритмы из имеющихся семейств и, используя соответствующие операции над алгоритмами (корректирующие операции), непосредственно строить из исходных алгоритмов оптимальный.

Одним из первых вариантов этой идеи явился так называемый корректор по результатам, предусматривавший формирование

решения задачи распознавания на основе результатов обработки исходной информации отдельными алгоритмами. Оказалось, однако, что не существует в некотором естественном смысле «хороших» простых операций, которые позволяли бы проводить необходимую коррекцию даже в том случае, когда в качестве допустимых ответов алгоритмов рассматриваются ответы «да», «нет», «не знаю». Дело в том, что пространство исходных информаций и множество возможных ответов определяются содержательной постановкой задачи. Поэтому первое состоит из достаточно сложно организованных элементов (обычно векторов очень большой размерности), а второе — весьма бедно ({0,1}).

В качестве выхода из этой ситуации нами были предложены способ определения алгоритма распознавания, в рамки которого укладываются все существующие типы алгоритмов, и так называемый алгебраический подход к задачам распознавания и классификации, обеспечивающий эффективное исследование и конструктивное описание классов алгоритмов распознавания. Этот подход предусматривает обогащение исходных эвристических семейств алгоритмов при помощи алгебраических операций и построение семейства, гарантирующего получение корректного алгоритма, обеспечивающего решение изучаемого класса задач.

В основе алгебраического подхода лежит идея индуктивного порождения математических объектов посредством обобщенного индуктивного определения. Выделяются базисные алгоритмы и модели распознавания и вводятся операции над ними, позволяющие последовательно порождать новые алгоритмы и модели. Выясняются условия, при которых данное семейство алгоритмов является базисным относительно введенных операций, а также свойства, которыми должна обладать модель, для того чтобы в ней нашелся алгоритм, правильно классифицирующий все объекты произвольной конечной выборки. Формируются методы построения таких алгоритмов. Смысл этого подхода состоит в том, что семейство таких алгоритмов рассматривается как некоторая алгебра, операции которой позволяют на основе базиса семейства алгоритмов строить такое расширение этого семейства, которое содержит корректный алгоритм, правильно классифицирующий конечную выборку по всем классам.

В алгебраическом подходе существенно используются особенности структуры, свойственные любой процедуре распознавания. Он предусматривает введение так называемого пространства оценок, промежуточного по отношению к исходным описаниям и допустимым ответам. Алгоритм распознавания при этом рассматривается как суперпозиция двух операторов. Первый из этих операторов — распознающий — в качестве ответов формирует элементы, называемые оценками, а второй (решающее правило) по оценкам определяет окончательные ответы. Таким образом, необходимость иметь дело с «неудобными» пространствами исходных описаний и допустимых ответов уступает место возможности вести

коррекцию в пространстве оценок (чаще всего оно представляет собой множество действительных чисел).

Важным в алгебраическом подходе является понятие полноты, связывающее отдельные задачи и модели алгоритмов: полнота некоторой задачи относительно модели означает, что при произвольном наборе априорных классификаций для рассматриваемых объектов в рамках модели может быть построен алгоритм, дающий всегда правильный ответ. Из полноты некоторой задачи относительно модели непосредственно следует существование в этой модели алгоритма, обеспечивающего абсолютную точность на материале обучения. Существенно, что построение экстремального алгоритма оказывается в большинстве случаев задачей, сравнительно легко разрешаемой стандартными математическими методами.

В рамках алгебраического подхода проведен ряд исследований, посвященных изучению и обоснованию развитых методов (некоторые из этих исследований нашли отражение в статьях данного ежегодника). Оказалось, что проблема границы множества корректирующих операций, переход за пределы которой в процессе расширения не дает реального эффекта, связана с фиксацией допустимого способа использования информации применяемыми алгоритмами. Формализация и последующее исследование содержательного представления о допустимом способе использования информации алгоритмами распознавания позволили получить ряд окончательных оценок для моделей алгоритмов и множеств корректирующих операций. Так, в частности, получена универсальная верхняя граница степени для множеств операций полиномиального типа, установлены нижние границы сложности для моделей распознающих операторов вычисления оценок и для моделей распознающих операторов, основанных на принципе разделения.

Показано, что возникающие при использовании алгебраического подхода семейства алгоритмов имеют ограниченную емкость и это обеспечивает корректность применения таких семейств в случае, когда выполнены некоторые достаточно общие гипотезы статистического характера. Оказалось, что во многих случаях экстремальные алгоритмы, формируемые в рамках алгебраического подхода, имеют ненулевой радиус устойчивости. Это означает, что при малом в некотором смысле изменении исходной информации классификация, порождаемая экстремальным алгоритмом, сохраняется, т. е. при выполнении достаточно общих допущений о компактности почти всюду имеет место сходимость классификаций, порождаемых экстремальными алгоритмами, к истинной. Проведены также исследования, связанные с изучением возможностей наиболее простого представления экстремальных алгоритмов.

Параллельно процессу перехода в распознавании от отдельных алгоритмов к моделям развивалась и другая ветвь исследований, связанных с использованием алгебраических методов для расширения типов исходных информаций, допустимых в задачах распознавания. В этой связи отметим теорию образов У. Гренандера

и развиваемую в рамках алгебраического подхода дескриптивную теорию анализа изображений, которая составляет основу нового научного направления в распознавании (см. также данный ежегодник).

Резюмируя изложенное, нам хотелось бы подчеркнуть, что методология распознавания используется в информатике в двух качествах:

— во-первых, по прямому назначению для решения задач распознавания в классическом смысле;

— во-вторых, как средство точного исследования плохо определенных задач.

В последнем случае эта методология реализовывается приблизительно следующим образом. Пусть, например, имеются некоторые данные, полученные в результате физического или имитационного эксперимента. Эти данные в некотором весьма ограниченном смысле характеризуют изучаемые объект или явление — необходимо попытаться свести их воедино, с тем чтобы установить, какие закономерности отражаются в имеющемся материале. Для этого выдвигается некоторая простая гипотеза, которой придается математический облик, и делается попытка «объяснить» имеющийся материал с ее помощью. Последовательное использование ряда эвристик (реализаций гипотезы) может позволить угадать модель. В противном случае происходит переход к поиску в рамках модели, порождаемой эвристикой, а затем к поиску оптимального (адекватного) эвристического принципа — модели. Если оказывается, что соответствующего принципа не существует или им нельзя практически воспользоваться, то следует сформировать некоторый конгломерат принципов, обеспечивающий выделение «федеративного» принципа,— именно этот верхний уровень и соответствует возможностям и назначению алгебраического подхода.