Введение в предмет распознавания образов. Кластерный анализ

- История проблемы
- Сущность проблем распознавания
- Развитие теории распознавания
- Область применения теории распознавания

Распознавание образов-объектов, сигналов, ситуаций, явлений или процессов представляет собой едва ли не самую распространенную задачу, с которой человеку приходится сталкиваться и решать ежечасно, ежеминутно, а порой и ежесекундно практически от первого до последнего дня своего существования. Для решения этой задачи человек использует огромные ресурсы своего мозга, включая одновременно, параллельно около 7-8 миллиардов нейронов. Именно это дает возможность людям практически мгновенно узнавать друг друга, с большой скоростью читать печатные и рукописные тексты – литературные, музыкальные, шахматные, безошибочно водить автомобили в сложном потоке уличного движения современного города, осуществлять отбраковку деталей на конвейере, дешифрировать аэро- и космические фотоснимки, разгадывать коды, древнюю египетскую клинопись и иероглифы народа майя. Распознавание представляет собой задачу преобразования входной информации, в качестве которой уместно рассматривать некоторые параметры, признаки распознаваемых образов в выходную, представляющую собой заключение о том, к какому классу относится распознаваемый образ. Именно поэтому, учитывая, что кибернетика есть наука об общих законах преобразования информации в сложных системах, распознавание образов представляет собой один из разделов этой науки.

В последнее время проблема автоматического распознавания образов была одной из наиболее актуальных и трудных проблем прикладной математики и математической кибернетики. Возникшая из решения конкретной задачи создания читающих автоматов, эта проблема в 60-е годы становится ведущей при попытках применять математику в слабо формализованных областях науки и практики, таких как практическая геология, биология, химия, медицина и т.п.

В течение достаточно продолжительного времени проблема распознавания привлекает внимание специалистов в области прикладной математики, кибернетики и информатики. Так можно, в частности, отметить работы Р. Фишера, выполненные в 20-х годах и приведшие к формированию дискриминантного анализа, как одного из разделов теории и практики распознавания. В 40-х годах А. Н. Колмогоровым и А. Я. Хинчиным поставлена задача о разделении смеси двух распределений.

Наиболее плодотворными явились 50-60-е годы XX века. В это время на основе массы работ появилась теория статистических решений. В результате этого появились алгоритмы, обеспечивающие отнесение нового объекта к одному из заданных классов, что явилось началом планомерного научного поиска и практических разработок. В рамках кибернетики начало формироваться новое научное направление, связанное с разработкой теоретических основ и практической реализации устройств, а затем и систем, предназначенных для распознавания объектов, явлений и процессов.

Многие исследователи считают, что задача классификации берет начало от работ Р. Розенблатта, который 1954 году выдвинул идею распознающего устройства порогового типа, предназначенного для перевода входных объектов в классы образов, называемого персептроном. Наряду с задачей обучения персептрона, поставлена задача самообучения, что дало основание бурному развитию теории “распознавание без учителя”.

Интенсивный рост статистической науки раскрыл новые возможности применения методов классификации в социально-экономических исследованиях, что в свою очередь,

стало поводом для активного исследования в области теории кластерного анализа. В этом направлении можно выделить работы Э. М. Бравермана, И. Б. Мучника, Л. А. Дорофеюка, Б. Г. Миркина, связанные с анализом матриц близости деревовидных структур.

Новая научная дисциплина получила название “Распознавание образов”.

Таким образом, базой для решения задач отнесения объектов к тому или иному классу послужили, как это отмечается сегодня, результаты классической теории статистических решений. В ее рамках строились алгоритмы, обеспечивающие на основе экспериментальных измерений параметров (признаков), характеризующих этот объект, а также некоторых априорных данных, описывающих классы, определение конкретного класса, к которому может быть отнесен распознаваемый объект.

В последующем математический аппарат теории распознавания расширился за счет применения: теории графов; теории информации; теории множеств; теории и методов статистического анализа; методов алгебры логики; информатики; математического программирования и системотехники.

Широкое распространение методов распознавания объясняется, во-первых, тем, что для их применения требуется значительно меньшая точность описания исследуемых объектов и явлений, чем при применении других математических методов.

Во-вторых, идея принятия решения на основе прецедентности является главной при формировании мировоззрения ученого-естественника. Действительно, обучение и исследование как продолжение обучения могут осуществляться лишь на базе изучения примеров. Примеры сопровождаются пояснениями, – почему этот пример требует тех или иных действий или почему он типичен для того или другого класса ситуаций. Тем самым мышление подготавливается к восприятию идеологии распознавания. Все допустимые объекты или явления разбиты на конечное число классов (классы могут пересекаться). Для каждого класса известно (ранее изучено) конечное число объектов (или явлений). Изучаемый новый объект следует, используя ранее накопленную информацию, отнести к тому или другому классу. Различные методы распознавания и являются формализацией описанной выше схемы.

Появление большого числа “внешних” для математики задач распознавания привело к возникновению двух школ с принципиально различной методикой исследования. Представители первой школы сделали попытку выделить в проблеме классы задач, в которых возможна формализация, достаточная для применения стандартных математических методов. Так возникли, например, статистические методы распознавания, созданные на базе принципов математической статистики. Основным недостатком работ этой школы было то, что требуемая степень формализации в реальных задачах обычно не выполнялась. Поэтому строгие математические результаты, получаемые в теоретических исследованиях, становились нестрогими в применениях.

Представители второй школы пришли к выводу, что появление новых типов задач требует создания принципиально новых подходов. Стало также ясно, что построение таких подходов требует проведения широких математических экспериментов. С появлением ЭВМ математики, как ранее естественники, получили возможность постановки экспериментальных исследований. Математический эксперимент по существу мало отличается от физического. Изучается реальная задача (ситуация), выдвигается гипотеза и затем проводится эксперимент, который подтверждает или опровергает гипотезу. Строгие математические доказательства на этом этапе не проводятся. В применении к задачам распознавания описанная выше традиционная схема выглядит так. Изучается класс реальных задач, приводящихся к схеме распознавания, например, задача прогнозирования в геологии. Изучение описаний месторождений и участков местности, где месторождения не

обнаружены, приводит к гипотезе: множества описаний первого и второго классов разделяются достаточно простой поверхностью.

Простейшей поверхностью является гиперплоскость. Уточненная гипотеза: описания, выполненные набором числовых признаков и принадлежащие разным классам, разделяются гиперплоскостью. Проводится эксперимент на ЭВМ и показывается, что в 99 случаях из 100 гипотетическое разделение действительно имеет место.

Появляется эвристический алгоритм. Если даны описания объектов двух разных классов, то следует построить поверхность, разделяющую эти описания. Для классификации нового объекта следует только установить, в какую часть пространства относительно выбранной поверхности попадает его описание.

Первый этап развития теории распознавания связан с построением большого числа существенно различных экспериментальных (эвристических) алгоритмов, которые в дальнейшем будем называть некорректными.

По мере накопления эвристических классифицирующих (некорректных) алгоритмов возникла необходимость в построении их единообразных описаний и теоретических обобщений. Детальный анализ таксономических процедур позволил описать общие принципы их формирования. Эти принципы, действующие уже над множествами алгоритмов, при некоторой формализации могут, реализовывались в виде математических описаний.

На втором этапе, имея в своем распоряжении богатый экспериментальный материал, математики приступают к изучению принципов формирования и строения некорректных процедур, хорошо решающих реальные задачи. Подобно физику-теоретику, выводящему свои уравнения на базе экспериментов, математик создает формальные описания классов распознающих процедур (модели распознающих алгоритмов) и проводит их исследование с помощью строгих математических методов.

Таким образом, появляются модели, построенные на принципе разделения, модели типа потенциалов, модели вычисления оценок и т.д.

Одной из таких моделей является модель алгоритмов распознавания образов, основанная на вычислении оценок (АВО), предложенная в начале 70-х годов Ю. И. Журавлевым и в дальнейшем развитая в его работах и в работах его учеников. Основы алгоритмов (АВО) составляет принцип вычисления оценок сходства, характеризующих близость между объектами, объектом и классами и между классами, по группе признаков, называемых опорными подмножествами заданного множества признаков. Очень часто полезная информация заключена не в отдельных признаках, а в различных их сочетаниях.

Исследование каждой из таких моделей имеет свои особенности, “внутреннюю” проблематику и проблематику, связанную с решением прикладных задач. Из последних наиболее важны задачи, связанные с отысканием в рамках модели оптимальных по точности алгоритмов для отдельных конкретных задач распознавания или классов задач. Построение таких оптимальных алгоритмов обычно приводится к исследованию, решению и созданию вычислительных схем для нестандартных экстремальных проблем.

На этой основе Ю. И. Журавлевым была поставлена задача построения общей теории алгоритмов распознавания и классификации и предложен так называемый “алгебраический подход к решению задач распознавания и классификации”, который позволяет проводить эффективное исследование и описание исследуемого класса алгоритмов.

Построение оптимального алгоритма в многопараметрической модели связано с решением трудных экстремальных задач. Предложенный алгебраический подход дает возможность для расширения семейства распознающих алгоритмов классификации с помощью специальных алгебраических операций и позволяет строить семейство алгоритмов, гарантирующих получение корректного алгоритма с заданными свойствами. Множество алгоритмов в этом случае является алгеброй, причем операции этой алгебры обладают

свойствами, позволяющими детально изучать различные характеристики исследуемого класса алгоритмов. Для описания класса алгоритмов, правильно классифицирующих исходную информацию, достаточно взять любую полную модель, рассмотреть линейное замыкание совокупности ее распознающих операторов и подключить к ней корректное решающее правило.

Таким образом, вместо построения формальных моделей в различных плохо формализованных областях достаточно построить семейство эвристических алгоритмов, затем ввести алгебру на множестве таких задач и построить его алгебраическое замыкание. В этом семействе оказывается принципиально разрешимой любая задача из множества плохо формализованных областей.

Широкий круг задач, возлагаемых на такие системы, определяется приведенным нами определением самого понятия “распознавания” и включает выяснение по разнородной, часто не полной, нечеткой, искаженной и косвенной информации факта, обладают ли изучаемые объекты, явления, процессы, ситуации фиксированным конечным набором свойств, позволяющим отнести их к определенному классу. Сюда входят как непосредственно задачи распознавания и классификации, так и такие задачи, в результате решения которых на основе распознавания требуется выяснить, в какой области из конечного числа областей будут находиться некоторые процессы через определенный промежуток времени.

Отсюда понятно, что к задачам распознавания должны относиться задачи технической и медицинской диагностики, геологического прогнозирования, прогнозирования свойств химических соединений, распознавания свойств динамических и статических объектов в сложной фоновой обстановке и при наличии активных и пассивных помех, прогнозирование урожая, обнаружения лесных пожаров, управления производственными процессами.

Начиная с 50-х годов разработано большое количество систем распознавания. Сегодня уже трудно назвать такую отрасль науки и сферы производства, где системы распознавания и классификации не используются или не будут использоваться в ближайшее время. При этом применение методов распознавания в ряде направлений науки и техники оказывает обратное, поистине революционизирующее влияние на эти направления. Рассмотрим некоторые области применения.

Системы технической диагностики. Их внедрение – важнейший фактор повышения эффективности использования машин и технологического оборудования, резкого сокращения расходов на эксплуатацию путем перехода к системам технической диагностики (распознавание состояния машин).

Медицинская диагностика. Автоматизированные системы диагностики в медицине – путь увеличения широты и глубины охвата симптомов, оперативности, достоверности.

Сельское хозяйство. Системы классификации и распознавания применяются для распознавания и прогнозирования размеров урожая по данным космических наблюдений, а также для уменьшения ручного труда при сортировке плодов по форме, цвету и размерам и т.п.