Таханов Р.С.

Что такое распознавание образов?

Первые исследования с вычислительной техникой в основном следовали классической схеме математического моделирования - математическая модель, алгоритм и расчет. Таковыми были задачи моделирования процессов происходящих при взрывах атомных бомб, расчета баллистических траекторий, экономических и прочих приложений. Однако помимо классических идей этого ряда возникали и методы основанные на совершенно иной природе, и как показывала практика решения некоторых задач, они зачастую давали лучший результат нежели решения, основанные на переусложненных математических моделях. Их идея заключалась в отказе от стремления создать исчерпывающую математическую модель изучаемого объекта(причем зачастую адекватные модели было практически невозможно построить), а вместо этого удовлетвориться ответом лишь на конкретные интересующие нас вопросы, причем эти ответы искать из общих для широкого класса задач соображений. К исследованиям такого рода относились распознавание зрительных образов, прогнозирование урожайности, уровня рек, задача различения нефтеносных и водоносных пластов по косвенным геофизическим данным и т. д. Конкретный ответ в этих задачах требовался в довольно простой форме, как например, принадлежность объекта одному из заранее фиксированных классов. А исходные данные этих задач, как правило, задавались в виде обрывочных сведений об изучаемых объектах, например в виде набора заранее расклассифицированных объектов. С математической точки зрения это означает, что распознавание образов(а так и был назван в нашей стране этот класс задач) представляет собой далеко идущее обобщение идеи эктраполяции функции.

Важность такой постановки для технических наук не вызывает никаких сомнений и уже это само по себе оправдывает многочисленные исследования в этой области. Однако задача распознавания образов имеет и более широкий аспект для естествознания(впрочем, было бы странно если нечто столь важное для искусственных кибернетических систем не имело бы значения для естесственных!). В контекст данной науки органично вошли и поставленные еще древними философами вопросы о природе нашего познания, нашей способности распознавать образы, закономерности, ситуации окружающего мира. В действительности, можно практически не сомневаться в том, что механизмы распознавания простейших образов, типа образов приближающегося опасного хищника или еды, сформировались значительно ранее чем возник элементарный язык и формально-логический аппарат. И не вызывает никаких сомнений, что такие механизмы достаточно развиты и у высших животных, которым так же в жизнедеятельности крайне

необходима способность различения достаточно сложной системы знаков природы. Таким образом, в природе мы видим, что феномен мышления и сознания явно базируется на способностях к распознаванию образов и дальнейший прогресс науки об интеллекте непосредственно связан с глубиной понимания фундаментальных законов распознавания. Понимая тот факт, что вышеперечисленные вопросы выходят далеко за рамки стандартного определения распознавания образов(в англоязычной литературе более распространен термин supervised learning), необходимо так же понимать, что они имеют глубокие связи с этим относительно узким(но все еще далеко неисчерпанным) направлением.

Уже сейчас распознавание образов плотно вошло в повседневную жизнь и является одним из самых насущных знаний современного инженера. В медицине распознавание образов помогает врачам ставить более точные диагнозы, на заводах оно используется для прогноза брака в партиях товаров. Системы биометрической идентификации личности в качестве своего алгоритмического ядра так же основаны на результатах этой дисциплины. Дальнейшее развитие исскуственного интеллекта, в частности проектирование компьютеров пятого поколения, способных к более непосредственному общению с человеком на естественных для людей языках и посредством речи, немыслимы без распознавания. Здесь рукой подать и до робототехники, исскуственных систем управления, содержащих в качестве жизненно важных подсистем системы распознавания.

Именно поэтому к развитию распознавания образов с самого начала было приковано немало внимания со стороны специалистов самого различного профиля - кибернетиков, нейрофизиологов, психологов, математиков, экономистов и т.д. Во многом именно по этой причине современное распознавание образов само питается идеями этих дисциплин. Не претендуя на полноту(а на нее в небольшом эссе претендовать невозможно) опишем историю распознавания образов, ключевые идеи.

История распознавания образов.

Рассмотрим кратко математический формализм распознавания образов. Объект в распознавании образов описывается совокупностью основных характеристик (признаков, свойств) $X = (x_1, ..., x_n)$, где i-я координата вектора X определяет значения i-й характеристики. Основные характеристики могут иметь различную природу: они могут браться из упорядоченного множества типа вещественной прямой, либо из дискретного множества(которое, впрочем, так же может быть наделено структурой). Такое понимание объекта согласуется как потребностью практических приложений распознавания образов, так и с нашим пониманием механизма восприятия объекта человеком. Действительно, мы полагаем, что при наблюдении(измерении)

объекта человеком, сведения о нем поступают по конечному числу сенсоров(анализируемых каналов) в мозг, и каждому сенсору можно сопоставить соответствующую характеристику объекта. Помимо признаков, соответствующих нашим измерениям объекта, существует так же выделенный признак, либо группа признаков, которые мы называем классифицирующими признаками, и в выяснении их значений при заданном векторе X и состоит задача, которую выполняют естественные и искусственные распознающие системы.

Понятно, что для того, чтобы установить значения этих признаков, необходимо иметь информацию о том как связаны известные признаки с классифицирующими. Информация об этой связи задается в форме прецедентов, то есть множества описаний объектов с известными значениями классифицирующих признаков. И по этой прецедентной информации и требуется построить решающее правило, которое будет ставить произвольному описанию объекта значения его классифицирующих признаков.

Такое понимание задачи распознавания образов утвердилось в науке начиная с 50-х годов прошлого века. И тогда же было замечено что такая постановка вовсе не является новой. С подобной формулировкой сталкивались и уже существовали вполне не плохо зарекомендовавшие себя методы статистического анализа данных, которые активно использовались для многих практических задач, таких как например, техническая диагностика. Поэтому первые шаги распознавания образов прошли под знаком статистического подхода, который и диктовал основную проблематику.

Статистический подход основывается на идее, что исходное пространство объектов представляет собой вероятностное пространство, а признаки(характеристики) объектов являют собой случайные величины заданные на нем. Тогда задача исследователя данных состояла в том, чтобы из некоторых соображений выдвинуть статистическую гипотезу о распределении признаков, а точнее о зависимости классифицирующих признаков от остальных. Статистическая гипотеза, как правило, представляла собой параметрически заданное множество функций распределения признаков. Типичной и классической статистической гипотезой является гипотеза о нормальности этого распределения (разновидностей таких гипотез статистики придумали великое множество). После формулировки гипотезы оставалось проверить эту гипотезу на прецедентных данных. Это проверка состояла в выборе некоторого распределения из первоначально заданного множества распределений(параметра гипотезы о распределении) и оценки надежности(доверительного интервала) этого выбора. Собственно эта функция распределения и была ответом к задаче, только объект классифицировался уже не однозначно, но с некоторыми вероятностями принадлежности к классам. Статистиками были

разработано так же и ассимптотическое обоснование таких методов. Такие обоснования делались по следующей схеме: устанавливался некоторый функционал качества выбора распределения (доверительный интервал) и показывалось, что при увеличении числа прецедентов, наш выбор с вероятностью стремящейся к 1 становился верным в смысле этого функционала(доверительный интервал стремился к 0). Забегая вперед скажем, что статистический взгляд на проблему распознавания оказался весьма плодотворным не только в смысле разработанных алгоритмов(в число которых входят методы кластерного, дискриминантного анализов, непараметрическая регрессия и т.д.), но и привел впоследствии Вапника к созданию глубокой статистической теории распознавания.

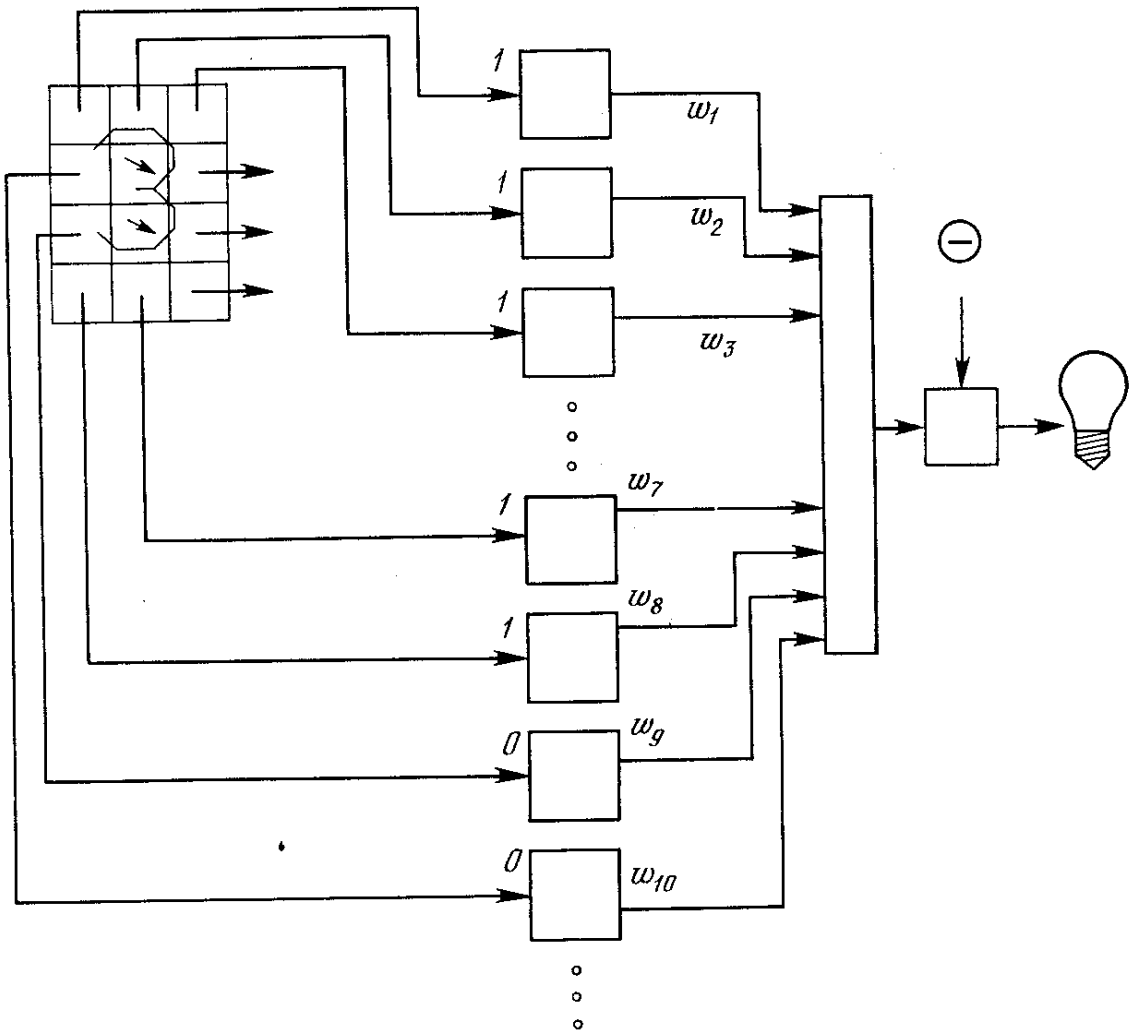
Тем не менее существует серьезная аргументация в пользу того, что задачи распознавания образов не сводятся к статистике. Любую такую задачу, в принципе, можно рассматривать со статистической точки зрения и результаты ее решения могут интерпретироваться статистически. Для этого необходимо лишь предположить, что пространство объектов задачи является вероятностным. Но с точки зрения инструментализма, критерием удачности статистической интерпретации некоторого метода распознавания может служить лишь наличие обоснавания этого метода на языке статистики как раздела математики. Под обоснаванием здесь понимается выработка основных требований к задаче которые обеспечивают успех в применении этого метода. Однако на данный момент для большей части методов распознавания, в том числе и для тех, которые напрямую возникли в рамках статистического подхода, подобных удовлетворительных обоснований не найдено. Кроме этого, наиболее часто применяемые на данный момент статистические алгоритмы, типа линейного дискриминанта Фишера, парзеновского окна, EM-алгоритма, метода ближайших соседей, не говоря уже о байесовских сетях доверия, имеют сильно выраженный эвристический характер и могут иметь интерпретации отличные от статистических. И наконец, ко всему вышесказанному следует добавить, что помимо асимптотического поведения методов распознавания, которое и является основным вопросом статистики, практика распознавания ставит вопросы вычислительной и структурной сложности методов, которые выводят далеко за рамки одной лишь теории вероятностей.

Итого, вопреки стремлениям статистиков рассматривать распознавание образов как раздел статистики, в практику и идеологию распознавания входили совершенно другие идеи. Одна из них была вызвана исследованиями в области распознавания зрительных образов и основана на следующей аналогии.

Как уже отмечалось, в повседневной жизни люди постоянно решают(зачастую бессознательно) проблемы распознавания различных ситуаций, слуховых и зрительных образов. Подобная

способность для ЭВМ представляет собой в лучшем случае дело будущего. Отсюда некоторыми пионерами распознавания образов был сделан вывод, что решение этих проблем на ЭВМ должно в общих чертах моделировать процессы человеческого мышления. Наиболее известной попыткой подойти к проблеме с этой стороны было знаменитое исследование Ф.Розенблатта по перцептронам.

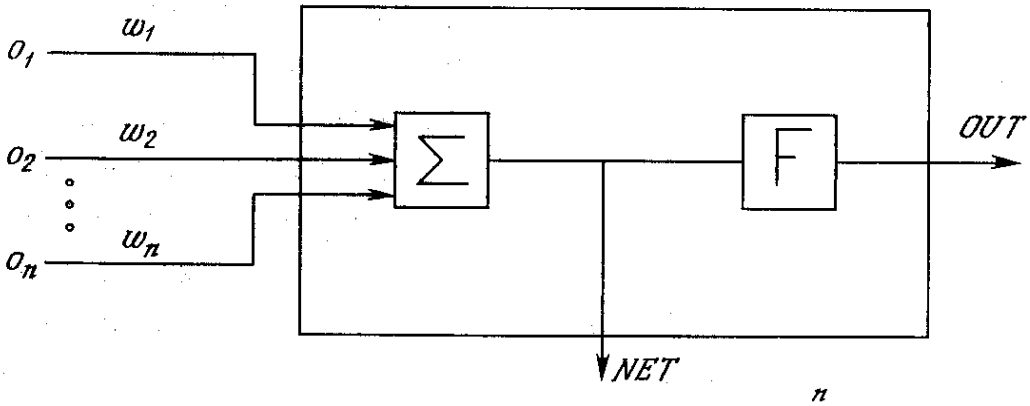
К середине 50-х годов казалось, что нейрофизиологами были поняты физические принципы работы мозга(в книге "Новый Разум Короля" знаменитый британский физик-теоретик Р. Пенроуз интересно ставит под сомнение нейросетевую модель мозга, обосновывая существенную роль в его функционировании квантово-механических эффектов; хотя, впрочем, эта модель подвергалась сомнению с самого начала. Отталкиваясь от этих открытий Ф.Розенблатт разработал модель обучения распознаванию зрительных образов, названную им персептроном. Персептрон Розенблатта представляет собой следующую функцию:



На входе персептрон получает вектор объекта $X=(x_1,...,x_n)$, который в работах Розенблатта предсталял собой бинарный вектор, показывавший какой из пикселов экрана зачернен изображением а

какой нет. Далее каждый из признаков подается на вход нейрона, действие которого на значение x_i представляет собой простое умножение на некоторый вес нейрона w_i. Результаты подаются на последний нейрон, который их складывает и общую сумму сравнивает с некоторым порогом w_0. В зависимости от результатов сравнения входной объект X признается нужным образом либо нет. Тогда задача обучения распознаванию образов состояла в таком подборе весов нейронов w_i и значения порога w_0, чтобы персептрон давал на прецедентных зрительных образах правильные ответы. Розенблатт полагал, что получившаяся функция будет неплохо распознавать нужный зрительный образ даже если входного объекта и не было среди прецедентов. Из бионических соображений им так же был придуман и метод подбора весов w_i и порога w_0, на котором останавливаться мы не будем. Скажем лишь, что его подход оказался успешным в ряде задач распознавания и породил собой целое направление исследований алгоритмов обучения основанных на нейронных сетях, частным случаем которых и является персептрон.

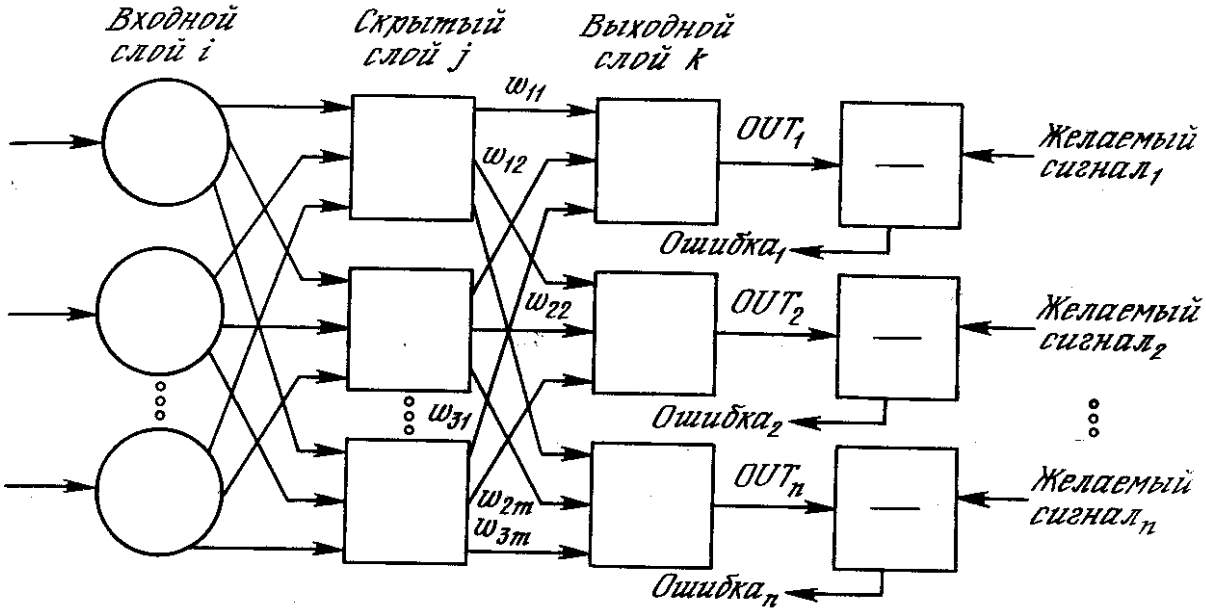
Далее были придуманы различные обобщения персептрона, функция нейронов была усложнена: нейроны теперь могли не только умножать входные числа или складывать их и сравнивать результат с порогами, но применять по отношению к ним более сложные функции. На следующем рисунке изображено одно из подобных усложнений нейрона:



$$NET = o_1 w_1 + o_2 w_2 + \ldots + o_n w_n = \sum_{i-1}^{n} o_i w_i$$

$$OUT = F(NET)$$

где $\mathrm{OUT} = \frac{1}{1+\mathrm{e}^{-\mathrm{NET}}}$. Кроме того топология нейронной сети могла быть значительно сложнее той, что рассматривал Розенблатт, например такой:



Усложнения приводили к увеличению числа настраиваемых параметров при обучении, но при этом увеличивали возможность настраиваться на очень сложные закономерности. Исследования в этой области сейчас идут по двум тесно связанным направлениям - изучаются и различные топологии сетей и различные методы настроек.

Нейронные сети на данный момент являются не только инструментом решения задач распознавания образов, но получили применение в исследованиях по ассоциативной памяти, сжатию изображений. Хотя это направление исследований и пересекается сильно с проблематикой распознавания образов, но представляет собой отдельный раздел кибернетики. Для распознавателя на данный момент, нейронные сети не более чем очень специфически определенное, параметрически заданное множество отображений, которое в этом смысле не имеет каких-либо существенных преимуществ над многими другим подобными моделями обучения которые далее будут кратко перечислены.

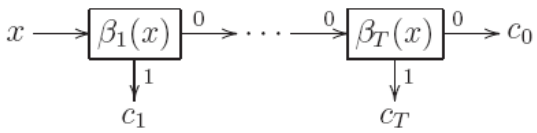
В связи с данной оценкой роли нейронных сетей для собственно распознавания(то есть не для бионики, для которой они имеют первостепенное значение уже сейчас) хотелось бы отметить следующее: нейронные сети, будучи чрезвычайно сложным объектом для математического анализа, при грамотном их использовании, позволяют находить весьма нетривиальные законы в данных. Их трудность для анализа, в общем случае, объясняется их сложной структурой и как следствие, практически неисчерпаемыми возможностями для обобщения самых различных закономерностей. Но эти достоинства, как это часто и бывает, являются источником потенциальных ошибок, возможности переобучения. Как будет рассказано далее, подобный двоякий взгляд на перспективы всякой модели обучения является одним из принципов машинного обучения.

Еще одним популярным направлением в распознавании являются логические правила и деревья решений. В сравнении с вышеупомянутыми методами распознавания эти методы наиболее активно используют идею выражения наших знаний о предметной области в виде, вероятно самых естественных(на сознательном уровне) структур - логических правил. Под элементарным логическим правилом подразумевается высказывание типа «если неклассифицируемые признаки находятся в соотношении X то классифицируемые находятся в соотношении Y». Примером такого правила в медицинской диагностике служит следующее: если возраст пациента выше 60 лет и ранее он перенёс инфаркт, то операцию не делать - риск отрицательного исхода велик.

Для поиска логических правил в данных необходимы 2 вещи: определить меру «информативности» правила и пространство правил. И задача поиска правил после этого превращается в задачу полного либо частичного перебора в пространстве правил с целью нахождения наиболее информативных из них. Определение информативности может быть введено самыми различными способами и мы не будем останавливаться на этом, считая что это тоже некоторый параметр модели. Пространство же поиска определяется стандартно. Предположим, что объект в нашей задаче характеризуется некласифицируемыми признаками $X=(x_1,...,x_i,...,x_n)$ и нужно установить значение классифицирумого признака y вне прецедентных данных. Тогда пространство состоит из правил вида «если $a_{i_1}<x_{i_1}<b_{i_1}$ и $a_{i_2}<x_{i_2}<b_{i_2}$ и ... и $a_{i_k}<x_{i_k}<b_{i_k}$, то $y=c_i$», причем число термов k слева ограничивается для того, чтобы избежать переобучения.

После нахождения достаточно информативных правил наступает фаза «сборки» правил в конечный классификатор. Не обсуждая глубоко проблемы которые здесь возникают(а их возникает немалое количество) перечислим 2 основных способа «сборки». Первый тип - линейный список. Пусть даны правила вида $\beta_i(x)\to c_i$, тогда результирующий классификатор определяется по принципу:

1: **для всех** $t=1,\dots,T$
2: **если** $\beta_t(x)$ **то**
3: **вернуть** c_t;
4: **вернуть** c_0.



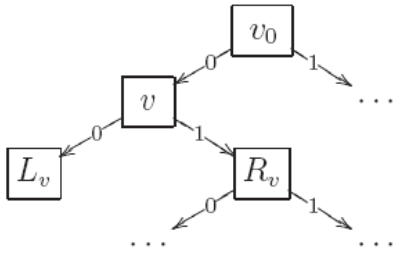
Второй тип – взвешенное голосование, когда каждому правилу ставится в соответствие некоторый вес, и объект относится классификатором к тому классу за который проголосовало наибольшее количество правил.

В действительности, этап построения правил и этап «сборки» выполняются сообща и, при построении взвешенного голосования либо списка, поиск правил на частях прецедентных данных

вызывается снова и снова, чтобы обеспечить лучшее согласование данных и модели.

Примыкают к логическим правилам и деревья классификации. В этом случае алгоритм классификации задается бинарным деревом, в котором каждой внутренней вершине v приписано предикат $\beta_v : X \to \{0,1\}$, каждой терминальной вершине v приписано имя класса c_v. При этом предикат β_v представляет собой терм вида $a_i < x_i < b_i$. При классификации объекта X, он проходит по дереву путь от корня до некоторого листа, в соответствии со следующим алгоритмом:

```
1: v := v_0;
2: пока вершина v внутренняя
3:     если β_v(x) = 1 то
4:         v := R_v;   (переход вправо)
5:     иначе
6:         v := L_v;   (переход влево)
7: вернуть c_v.
```



Фактически, бинарные деревья классификации выражают идею построения промежуточных "концептов", которым соответствуют вершины дерева. И тогда классификация представляет собой логический вывод по правилу modes ponens, исходной аксиомой которого является корень дерева, а дедуктивными предположениями - утверждения вида "если объект опустился до вершины v и удовлетворяет предикату этой вершины β_v, то опустить его в R_v " и "если объект опустился до вершины v и не удовлетворяет предикату этой вершины β_v, то опустить его в L_v". И тогда обратный процесс - построения этих концептов и правил можно интерпретировать как индуктивный вывод(inductive learning). Понятно, что предположение о древесной структуре такого логического вывода есть следствие чрезвычайной сложности вывода в общем случае, хотя, вероятно, другие системы вывода еще будут открыты.

На этом закончим наше краткое ознакомление с современными методами распознавания. Я полагаю, что общее представление о том, какие методы сейчас в моде, у читателя уже появились. Поскольку даже краткий обзор всех ныне существующих подходов к задачам распознавания нереально провести в пределах маленького эссе, мы лишь перечислим некоторые другие методы, снабдив необходимыми ссылками: байесовские сети доверия, методы опорных векторов Вапника, потенциальных функций, алгоритмы вычисления оценок(АВО), метрические алгоритмы.

Теория Вапника-Червоненкиса

Заметим, что обоснование методов распознавания является характерной особенностью именно статистического подхода. Это и неудивительно - никакая другая формализация распознавания не

имеет столь глубоко разработанную математическую основу как статистический подход, базой которого служит теория вероятности. И хотя попытки создать математическую основу, отличную от вероятностной, для оценок качества алгоритмов предпринимаются, ни одну из них успешной назвать пока нельзя.

Что вообще в данном случае подразумевается под обоснованием?

Чтобы понять это рассмотрим модельную(и классическую) задачу распознавания образов. Пусть в пространстве R^n заданы конечное число объектов 2 классов: A и B. Объекты классов A и B появляются с некоторыми неизвестными вероятностями p_A и p_B соответственно, причем распределения объектов A и B подчиняются нормальному закону. И требуется построить решающее правило, которое разделяло бы пространство R^n на 2 подмножества - объекты класса A и B, и причем с малой вероятностью ошибалась бы на внешних данных. Эта задача была рассмотрена в 1936 г. Рональдом Фишером, который и разработал теорию ее решения. Фишер предложил следующее решение: вначале оценить по обучающей выборке вероятности p_A и p_B, а затем оценить параметры нормальных законов элементов из A и B(то есть оценить соответственно (μ_A,Σ_A) и (μ_B,Σ_B); здесь первый элемент в паре матожидание, а второй - ковариационная матрица, нормальные законы обозначим как $N(x,\mu_A,\Sigma_A)$ и $N(x,\mu_B,\Sigma_B)$). Далее решающее правило имело интуитивно очевидный вид: объект классифицировался как A, если $p_A N(x,\mu_A,\Sigma_A) > p_B N(x,\mu_B,\Sigma_B)$, и как B, если $p_A N(x,\mu_A,\Sigma_A) < p_B N(x,\mu_B,\Sigma_B)$.

Обосновывался этот алгоритм довольно просто. При достаточно большой величине обучающей выборки вероятности p_A и p_B оцениваются с большой точностью по закону больших чисел, а параметры (μ_A,Σ_A) и (μ_B,Σ_B) оценивались соответственно по частям выборки класса A и B с помощью стандартных статистик, доверительные интервалы которых, как известно, стремятся к 0 при увеличении числа точек по которым они оцениваются. И наконец, в случае, когда нормальные распределения (μ_A,Σ_A) и (μ_B,Σ_B) достаточно хорошо отстоят относительно друг друга(если они перекрываются на 99%, хорошей обобщающей способности ждать не приходится), то можно с большой вероятностью(которая может быть оценена) ожидать правильной классификации вне обучения.

Итого, для того, чтобы данный алгоритм хорошо обобщал требуется чтобы на обучении было достаточно много объектов как класса A так и B, и эти классы были достаточно отлично друг от друга распределены. Этим качественным соображениям соответствует достаточно хорошо проработанная количественная статистическая теория оценок.

Таким образом, под обоснованием метода распознавания будем считать некоторый набор количественных критериев, удовлетворение которых обеспечивает с разумной вероятностью хорошую обобщающую способность алгоритма.

В конце 60 - начале 70 годов В.Н.Вапником и А.Я.Червоненкисом (далее везде VC) была создана статистическая теория распознавания, которая несмотря на некоторые свои недостатки, и стала основным инструментом в обосновании методов распознавания.

VC исходили из следующей модели обучения: предполагалось, что фиксировано множество функций $F \subset \{f : O \to C\}$, O - множество классифицируемых объектов, и $C = \{0,1\}$ - множество классов (при фиксировнном некотором классе c, это множество можно интерпретировать как принадлежность к c и непринадлежность к c). Тогда при предъявлении методу обучающей выборки $\{(x_i, y_i)\}_{i=1}^{l}$ (элементы которой последовательно и независимо выбираются из множества $O \times C$ согласно некоторому неизвестному распределению ρ; обозначим эту выборку X^l), метод выбирает некоторую функцию $f \in F$, которая и является результирующим классификатором. Теория VC занимается тем, что оценивает вероятность того факта, что ошибка на контроле существенно превысит ошибку на обучении, то есть функционал $P\left(\nu\left(\mu\left(X^l\right), X^k\right) > \nu\left(\mu\left(X^l\right), X^l\right) + \varepsilon\right)$ (здесь X^k - контрольная выборка, которая также независимо выбирается по ρ), причем эта оценка получается для любого распределения на $O \times C$ (то есть это оценка в худшем случае).

Выглядит она следующим образом:
$P\left(\nu\left(\mu\left(X^l\right), X^k\right) > \nu\left(\mu\left(X^l\right), X^l\right) + \varepsilon\right) \le \Delta^F(2l) \cdot 1{,}5e^{-\varepsilon^2 l}$, где $k = l$ (будем полагать для простоты, что длина контроля равна длине обучения).

В правой части стоят 2 сомножителя - второй $1{,}5e^{-\varepsilon^2 l}$ экспоненциально стремится к 0, при длине обучающей выборки стремящейся к бесконечности. Интерес представляет первый сомножитель, называемый сложностным. Для любого подмножества M^{2l} множества объектов мощности $2l$ введем $S^F\left(M^{2l}\right)$ для обозначения множества различных разбиений M^{2l} на классы, индуцируемые функциями F. Тогда, по определению, $\Delta^F(2l) = \max_{M^{2l}} S^F\left(M^{2l}\right)$. Таким образом, $\Delta^F(2l)$ характеризует разнообразие множества функций F ограниченных на множества длины $2l$. Таким образом, $\Delta^F(2l)$ зависит только от множества функций откуда берется результирующее отображение и не зависит от алгоритма который используется для получения его. И если предположить, что $\Delta^F(2l)$ при $2l \to \infty$ растет

медленнее любой экспоненты $e^{\alpha l}$, то можно утверждать, что $P\left(\nu\left(\mu\left(X^l\right),X^k\right)>\nu\left(\mu\left(X^l\right),X^l\right)+\varepsilon\right)\to 0$, при $k=l\to\infty$, что и обеспечивает малое отклонение ошибки на контроле от ошибки на обучении.

VC нашли так же и важный частный случай, когда $\Delta^F\left(2l\right)$, при $2l\to\infty$, растет медленнее любой экспоненты. Они ввели важную структурную характеристику множества F называемую VC-размерностью.

VC-размерность определяется следующим образом. Это максимальное число d, такое, что найдется набор из d классифицируемых объектов, который может быть классифицирован функциями из F всеми 2^d способами. Если F имеет конечную VC-размерность d, то можно показать, что $\Delta^F\left(2l\right)\le\left(\frac{2l}{d}\right)^d$, то есть $\Delta^F\left(2l\right)$, как и требуется, растет достаточно медленно. И отсюда следует, что конечная VC-размерность обеспечивает асимптотически малую разницу между ошибками на обучении и контроле.

Дальнейшее развитие теории Вапника-Червоненкиса заключалось доказательстве для многих методов обучения конечности VC-размерности. Выяснилось, что для параметрических моделей обучения, как правило, VC-размерность имеет порядок числа параметров необходимых для настройки модели.

В качестве иллюстративного примера вернемся к задаче которая обсуждалась вначале. Фишером, помимо вышеуказанного алгоритма, была придумана и интересная эвристика - он заметил, что при одинаковых ковариационных матрицах Σ_A и Σ_B, оптимальное решающее правило вырождается в линейный тип: если $w_0+w_1x_1+\ldots+w_nx_n>0$ то ответ 1, иначе 0, при некоторых весах w_i. Тогда он предложил использовать эту эвристику даже в случаях, когда ковариационные матрицы не равны, и тем самым им впервые была применена эвристика, основанная на линейной разделимости!

Но уже для эвристического алгоритма старое обоснование не работало. Однако оно хорошо ложится на канву теории VC. Ясно, что в данном случае $F=\left\{f:R^n\to\{0,1\}\mid\exists\overline{w}f\left(\overline{x}\right)=\left[w_0+w_1x_1+\ldots+w_nx_n>0\right]\right\}$. Легко видеть, что никакие $n+2$ точки не могут быть разделены линейными гиперплоскостями в R^n всеми способами, и потому, VC-размерность конечна. Отсюда можно заключить, что ошибка на обучении ассимптотически сколь угодно близка к ошибке на контроле и это доказывает тот факт, что разделимость большого числа случайно взятых точек в n-мерном пространстве не может быть случайным фактом (что, в общем-то, очевидно) и влечет из себя хорошую обобщающую способность вне обучения!

Интерпретация оценки VC имеет связь с идущем со средневековья принципом Оккама – “не употребляй сущностей без

надобности". Первый сомножитель в оценке, $\Delta^F(2l)$, неслучайно был назван сложностным. Как уже замечалось он характеризует разнообразие множества функций F ограниченных на множества длины $2l$. И добавочная ошибка(то есть ошибка прибавляемая к ошибке на обучении, чтобы получить верхнюю VC-оценку ошибки на контроле) пропорциональна логарифму сложностного сомножителя. Из этого факта следует, что чем меньше сложностный сомножитель, тем меньше эта добавка. Но с другой стороны, если сложность множества F мала, то мы не сможем хорошо выбрать $f \in F$ с малой ошидкой на обучающей выборке. Итого, мы имеем 2 противоположных эффекта от увеличения сложностного сомножителя, положительный эффект подгона под обучение и отрицательный эффект увеличения сложности, который влечет собой переобучение(overfitting). Поиск равновесия между ними и является математической формой(пусть упрощенной) принципа Оккама: нужно выбирать из всех методов, обеспечивающих данную ошибку на обучении, самый простой(в смысле сложностного члена).

Однако теория VC, несмотря на то, что ассимптотически хорошо объясняла многие методы распознавания, давала оценки необходимых длин обучения порядка $10^6 - 10^8$, которые оказались совершенно несовместимыми с реальными достижениями методов обучения, дающими результаты при сотнях объектов(например тот же алгоритм поиска линейной разделяющей гиперплоскости). Существующий зазор между теорией и практикой распознавания до сих пор не заполнен. И хотя такая ситуация, в принципе, характерна для современного состояния вещей в computer science, видимо, полное решение этой проблемы, только для ныне существующих методов обучения, до какой-то степени завершит целый этап в распознавании образов.

Почему же оценки VC оказались столь завышенными? Вероятно, самым простым объяснением является чрезмерная их общность: они не учитывают распределения на множестве объектов, а способ выбора функции $f \in F$ по обучающей выборке для них является черным ящиком.

На идее принципа Оккама, кстати говоря, был основан алгоритм структурной минимизации риска. В нем начале фиксировались расширяющиеся множества функций $F_1 \subset F_2 \subset \ldots$. Он работал по принципу: на множестве F_i обучалась функция $f \in F_i$, которая допускала долю ν_i ошибок на обучении, и тогда по теории VC ошибка на контроле оценивалась как $\nu_i + \varepsilon(F_i, l)$, где $\varepsilon(F_i, l)$ - сложностная надбавка. И находилось наилучшее F_i по критерию $\nu_i + \varepsilon(F_i, l)$. В данном случае интересно, что данный алгоритм берет функцию из

множества $\bigcup_i F_i$, которое, вообще говоря, может иметь сколь угодно большую емкость и сложность!

Уже отсюда видно, вышесформулированный математический принцип Оккама не вполне корректен – очень трудно сложность описать одним скаляром! Одним из главных аспектов сложного является то, что понимать его можно различно: вселенная, как сказал некто, подобна сыру, который может быть разрезан большим числом способов, а разрез выбирается вами из соображений удобств. И сложность так же многолика.

Композиции методов распознавания.

Как следует из теории VC, если в низкоразмерном пространстве разделена куча объектов, мы понимаем, что это «неслучайно» и должны радоваться. Но вероятность такой «неслучайности» мала и объекты могут не разделиться. Но мы не унываем и полагаем, что раз в данных нет закономерности типа линейной разделимости, то могут быть и другие! Мы применяем другой алгоритм распознавания и т.д. до тех пор, пока не получим результат.

Таким образом, если у нас нет априорных знаний о виде, в котором нужно искать закономерности, ничего не остается как пробовать различные модели – метрические, основанные на логических правилах и т.д.

Если посмотреть на данную схему с точки зрения теории VC, то можно заметить, что функция роста $\Delta^F(2l)$ такой «метамодели» будет суммой функций роста каждой из моделей участвующих в ней. Такое относительно незначительное усложнение может тем не менее серьезно улучшить способность удовлетворить ограничения обучающей выборки, при отсутствии априорных знаний о данных.

Однако вполне возможен и такой вариант, когда какие-то законы в данных ловятся посредством одной модели, а какие-то посредством другой.

Предположим, например, что задачей системы распознавания является прогноз динамики некоторого объекта по наблюдению его истории. И пусть этот объект может находится в некоторых "относительно хороших" состояниях $S_1,...,S_n$, в каждом из которых его поведение может быть хорошо спрогнозировано "относительно простыми" методами распознавания $M_1,...,M_n$. В этом случае имеет смысл строить систему распознавания каким-то образом комбинируя $M_1,...,M_n$, например, с весами характеризующими наше знание о том, в каком из состояний $S_1,...,S_n$ находится система(здесь еще неявно зашито предположение, что есть "относительно простая" система распознавания прогнозирующая в каком из $S_1,...,S_n$ находится объект). Тогда подобное комбинирование помогло бы настроится на сложное поведение не переусложненным методом.

Отсюда возникла идея композиции методов распознавания. В СССР для решения подобных задач Ю.И. Журавлевым был разработан алгебраический подход. Поскольку алгебраический подход находится на острие современной теории распознавания, строгих теоретических результатов в этой области не так много.

Математический формализм который возникает при анализе подобных алгоритмов весьма сложен. Это и неудивительно, сложно получить объяснения работы ансамбля сложно взаимодействующих алгоритмов, если до сих пор непонятно, почему может работать один.

Однако уже сейчас понятно, что одним из основных вопросов интересующих теорию композиций алгоритмов являются условия при которых соответствующие модели смогут обеспечить полное удовлетворение ограничений обучающей выборки – так называемые условия разрешимости. На этой идее основывается весьма глубокая теория локальных и универсальных ограничений выдвинутая К.В. Рудаковым.

Статистическое рассмотрение конструкций алгебраического подхода показало, что ассимптотическое обоснование(а других пока от статистики и не дождешься) для некоторых композитных алгоритмов, типа бустинга, алгебраических расширений АВО, так же может быть найдено.

Закончить это эссе хочется следующим замечанием. Первые системы распознавания, которые были основаны на статистике, включали в качестве предварительных к собственно построению решающего правила немало шагов по подготовке данных. В качестве обязательной части системы распознавания выступали люди, которые совершали предварительные исследования данных, поиски самых очевидных законов, всевозможные «озарения», на основе которых принимали решения какую же модель распознавания использовать. Если одним из результатов развития направления связанного с композициями алгоритмов будет хотя бы частичное снятие необходимости в подобной предварительной подготовке данных человеком, то это направление распознавания образов уже выполнит свое важнейшее назначение.

1.Rosenblatt, F. (1962). Principles of Neurodynamics. Spartan Books.
2.Fisher, R.A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7: 179-188 (1936)
3.J. R. Quinlan, “Induction of decision trees,” Mach. Learn., vol. 1, pp. 81–106, 1986.
4.E. Parzen, On Estimation of a Probability Density Function and Mode, Annals of Math. Statistics, 33: 1065–1076, 1962.
5.Arthur Dempster, Nan Laird, and Donald Rubin. "Maximum likelihood from incomplete data via the EM algorithm". Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977

6.B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, 1992. ACM Press.
7.Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1999. ISBN 0-387-98780-0
8.Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проб. кибернетики. М.: Наука, 1978. Вып. 33. С. 5-68
9.Рудаков К.В. Об алгебраической теории универсальных и локальных ограничений для задач классификации // Распознавание, классификация, прогноз. Математические методы и их применение. Вып. 1. - М.: Наука, 1989. - С. 176-200.
10.Растригин Л. А., Эренштейн Р. Х. Метод коллективного распознавания. 79 с. ил. 20 см., М. Энергоиздат 1981
11.Мазуров В.Д. Комитеты систем неравенств и задача распознавания // Кибернетика, 1971, № 2. С. 140-146.
12.Минский М., Пейперт С. Персептроны. - М.: Мир, 1971.
13.Фу К. Структурные методы в распознавании образов. - М.: Мир, 1977.
14.Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. - М.: Наука, 1970.
15.Журавлев Ю.И. Избранные научные труды. – Изд. Магистр, 1999.
16.Pearl J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann Publishers, 552 pp.