АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И ТАБЛИЦ ДАННЫХ
(Вычислительные системы)
1994 год Выпуск 150

УДК 519.95

АНАЛИЗ ДАННЫХ И АНАЛИЗ ЗНАНИЙ

Н.Г. Загоруйко

Определяется место методов анализа данных среди других направлений прикладной математики. По аналогии с задачами анализа данных формулируются задачи анализа знаний. Рассматриваются возможные методы решения этих задач с использованием меры близости в пространстве знаний.

1. Задачи анализа данных

Вначале поясним место среди задач прикладной математики того направления, которое с подачи французских математиков получило название "анализ данных".

Классическое направление прикладной математики связано с методами вычислений одних характеристик изучаемого объекта или явления по известным значениям других его характеристик. При этом модель объекта считается известной и зависимости между характеристиками описываются аналитическим выражением в виде уравнения или системы уравнений или неравенств. Проблемы, возникающие при решении таких задач, связаны с большими объемами вычислений, с защитой от погрешностей, накапливающихся в компьютере из-за округления чисел и т.д.

Позже появились задачи анализа объектов, математическая модель которых известна с точностью до параметров.Известен на-

бор характеристик, влияющих на целевую характеристику, извес - тен также общий вид зависимости между характеристиками,но коэффициенты, показатели степени и другие параметры модели неизвестны и, чтобы их определить, используются протоколы наблюдений, отражающие значения одних характеристик при разных значениях других. Делается серия предположений о значениях неизвестных параметров модели и эти предположения проверяются на про - токолах. В результате выбираются такие значения параметров,при которых модель с заданной точностью позволяет по одним (входным) характеристикам определять другие (выходные или целевые) характеристики. Такого рода задачи называются задачами "иден - тификации моделей".

Наконец, с появлением кибернетики стали формулироваться задачи анализа "черного ящика": исследователю известен набор характеристик, среди которых имеются характеристики, влияющие на целевое свойство объекта, но какие из них являются опреде - ляющими ("информативными") и какой математической моделью описываются закономерности их влияния на целевую характеристику - неизвестно. Нужно выбрать информативные характеристики и по - строить модель, позволяющую вычислять значения одних характе - ристик по значениям других. Единственным источником информации для решения такой задачи служит таблица экспериментальных данных типа "стимул - реакция" с описанием входных и выходных характеристик наблюдаемого объекта или множества объектов. Часто такие таблицы данных называют таблицами "объект - свойство".Теперь выбор класса моделей и конкретной модели с определенными параметрами сверяется с материалом таблицы данных. Возникающий при этом круг задач и составляет направление, именуемое зада - чами "анализа данных".

Возвращаясь к началу,можно отметить, что вычислительная математика обычно не имеет дела с этапом выдвижения гипотез о том,какие характеристики должны быть включены в модель объекта

и какой должна быть эта модель. Риск сделать ошибку в выборе модели и ее параметров остается вне поля внимания и аккуратные вычисления по имеющейся модели создают впечатление высокого качества решения проблемы в целом.

Задачи идентификации моделей требуют от математика ответственности за правильный выбор параметров модели. Наличие этого рискованного шага в процессе решения задачи лишает результат ореола строгой математической чистоты.

На результатах решения задач анализа данных лежит явный след большого числа рискованных предположений: и о выборе характеристик объекта, и о классе моделей, и о параметрах выбранной модели. Эти предположения представляются на языке математических формул, но природа их появления лежит вне математики, так что основная часть процесса решения задач анализа данных связана с проникновением в природу изучаемого явления и характерна скорее для естественно-научных областей. Ситуация усугубляется еще и тем, что реальные данные обладают такими особенностями, которые затрудняют применение строгих математических методов. Достаточно отметить, что таблицы данных часто бывают представлены малыми выборками в пространствах большой размерности при отсутствии информации о характере и степени зависимости одних характеристик от других, при разнотипности измерительных шкал, наличии шумов и пробелов. В этих условиях методы решения задач анализа данных вынужденно основываются как на корректных математических процедурах, так и на чисто эвристических приемах. Не удивительно, что получаемые решения воспринимаются настороженно, а многие методы решения выглядят недостаточно строго обоснованными.

Это обстоятельство объективно отражает тот факт, что на любом этапе развития прикладной математики возникают реальные задачи, для решения которых еще не готовы хорошо обоснованные

математические методы. Вместе с тем, важность задач не позволяет отложить их решение и вынуждает принимать рискованные эмпирические гипотезы и использовать нестрогие эвристические приемы. Если получаемые при этом результаты (предсказания,прогнозы) подтверждаются фактами, то настороженность в восприятии использованной модели сменяется уверенностью в ее адекватности изучаемому явлению и внимание математика переносится на аналитическое исследование модели и на вычислительные трудности,связанные с ее использованием.

А доброжелательные и стимулирующие термины типа "голая эвристика", "мутный поток литературы", "бред сивой кобылы" применяются уже к попыткам решения других задач в кипящем слое новых прикладных проблем.

После такого отступления, навеянного многолетней практикой участия в дискуссиях между "чистыми" и "прикладными" математиками, можно обратиться непосредственно к задачам анализа данных, которые, как кажется, уже прошли значительную часть пути от первых постановок до адекватных моделей.

Достаточно подробная классификация задач анализа данных приведена в работе [1]. Среди них можно выделить несколько задач, которые встречаются в практике анализа данных наиболее часто и изучены лучше других [2,3].

Прежде всего, это задачи предсказания одного, нескольких или всех элементов некоторого (целевого) признака, измеренного в шкале наименований. Если предсказать нужно один элемент целевого классификационного признака, то мы имеем дело с хорошо изученной задачей распознавания образов.

Если нужно предсказать (сформировать) все элементы этого признака, то решается задача таксономии или автоматической классификации. Предсказание нескольких элементов классификационного признака занимает промежуточное значение и решается с помощью таксономических решающих функций.

С предсказанием элементов строки связана важная задача выбора подсистемы информативных признаков.

Если требуется предсказывать произвольное множество элементов, принадлежащих разным строкам и столбцам таблицы данных, то мы имеем дело с задачей заполнения пробелов.

2. Метрика в пространстве знаний

Информация, которая используется в экспертных системах часто бывает представлена в виде продукций типа "Если X_1 & X_2 & ..., то A". При этом значения переменных могут задаваться разным способом, например: $X_1 = 7$; $X_2 = (2\text{-}6)$; $X_3 = a \vee b \vee c$; $X_4 > 0$ и т.д. Такой специфичный вид представления знаний налагает большие ограничения на методы работы с ними. Методы логического вывода, опирающиеся на сравнение левых и правых частей двух продукций, рассматривают все переменные через призму шкалы наименований [4], и результат сравнения считается положительным, если имеет место точное совпадение значений сравниваемых предикатов. Величина отличия значений предикатов роли не играет, номинальная шкала не позволяет оперировать такими понятиями как степень "непохожести", "близости", "аналогичности", т.е. понятиями, на которых основаны человеческие способы рассуждений по аналогии. Ясно, что для расширения логических возможностей экспертных систем нужно научиться измерять степень "похожести" знаний или ввести метрику для измерения расстояний в пространстве знаний.

Такая метрика была введена в [5]. Можно считать, что каждый предикат отражает знание эксперта о распределении возможных значений данной характеристики. Утверждение $X_3 = (a \vee b \vee c)$ равносильно утверждению, что предикат X_3 с одинаковой вероятностью (1/3) может принимать одно из трех значений, а условие $X_2 = (2\text{-}6)$ означает, что значение предиката X_2 с вероятностью 0,25 может принять одно из четырех значений в диапазоне от 2

до 6. Следовательно, расстояние между одноименными предикатами можно определять через расстояние между двумя распределениями вероятностей. Была сконструирована мера для измерения этого расстояния $R = f(r, h, w)$, которая учитывает расстояние r от всех элементов одного распределения до всех элементов другого, энтропийную меру h, близкую по смыслу к дисперсии распределений, и степень w пересечения распределений (величину области "консенсуса").

Эти аргументы вычисляются так: разделим ось X, отображающую мнение первого эксперта о распределении предиката P, на m частей ("квантилей") так, чтобы в каждой части была заключена плотность вероятности, равная 1/m. Правая граница первого квантиля находится в точке $X_{1,1}$, второго - в точке $X_{1,2}$, i-го - в точке $X_{1,i}$ и т.д. до $X_{1,m}$. Аналогично, границы квантилей распределения, указанного вторым экспертом, будут находится в точках $X_{2,1}$, $X_{2,2}, \ldots, X_{2,i}, \ldots, X_{2,m}$. Расстояние r будем определять так*):

$$r = \sum_{i=1}^{m} |X_{1,i} - X_{2,i}|.$$

Принимается, что область пересечения w равна 0, если два распределения не пересекаются, и равна 1, если распределения совпадают полностью. В промежуточных случаях

$$S = \sum_{q=1}^{Q} |P_{1,q} - P_{2,q}| * 0{,}5,$$

если Q - число делений, равномерно распределенных вдоль оси X. Чем больше область консенсуса, тем меньше расстояние между значениями, следовательно, расстояние R должно быть пропорциональным величине $w = (1 - S)$.

Расстояние между суждениями экспертов зависит и от категоричности их оценок. При одном и том же расстоянии r мера R счи-

*) Автор благодарен В.В.Бовнеру за полезные обсуждения этого раздела работы.

тается тем большей,чем меньше энтропия h в суждениях экспертов. Величина h находится следующим образом: $h = (h_{max} - h_{1,2})/h_{max}$, где $h_{max} = \ln m$, $h_{1,2} = 0{,}5 * (h_1 + h_2)$. Здесь

$$h_1 = \sum_{q=1}^{Q} P_{1,q} * \ln P_{1,q}, \quad h_2 = \sum_{q=1}^{Q} P_{2,q} * \ln P_{2,q}.$$

Общая мера расстояния между двумя знаниями о характеристике X теперь принимается равной

$$R = r * w * h. \qquad (1)$$

Эта мера удовлетворяет таким естественным аксиомам как непрерывность,симметричность и транзитивность.В [5] приводятся способы вычисления меры R для порядковых и номинальных шкал.

Если эксперт не высказывается о значении некоторой характеристики, то это означает, что он либо не знает этого значения, либо считает данную характеристику несущественной. В таком случае можно считать, что для него все значения характеристики равновероятны. Это предположение позволяет находить расстояние между знаниями, если даже эксперты оперируют не полностью совпадающими наборами характеристик.

Проверка правомочности применения описанной меры делалась путем экспертного оценивания. Были предъявлены различные пары распределений и эксперты упорядочивали эти пары по степени их "похожести", "близости". Мера близости, найденная по приведенной формуле, сохраняла установленный экспертами порядок.

Появление меры близости в пространстве знаний позволило реализовать в экспертной системе партнерского типа ЭКСНА логический вывод, использующий рассуждения по аналогии [6,7]. На этой же мере основан в ЭКСНЕ блок автоматического обнаружения противоречий между знаниями, что особенно важно в процессе отладки базы знаний и при пополнении базы знаний в ходе эксплуатации системы: если расстояние между условиями в продукциях (т.е. между их левыми частями) малы, а расстояния между следствиями (правыми частями) велики, то это значит, что из одних

и тех же условий вытекают разные следствия, что указывает на противоречия между двумя знаниями.

3. Задачи анализа знаний

При решении таких задач анализа данных как таксономия выбор информативной подсистемы признаков, распознавание образов и заполнение пробелов в таблицах данных существенно использу - ются меры расстояний между объектами или между признаками [8]. После появления меры для измерения расстояний в пространстве знаний появилось естественное желание сформулировать аналоги указанных выше задач, но применительно к знаниям. Что собой представляет задача таксономии знаний, как ее можно было бы решать и для чего применять? Какова содержательная интерпретация задачи, аналогичной задаче выбора информативной подсистемы признаков? Как можно было бы задать описание образа в простран - стве знаний и построить решающую функцию для распознавания принадлежности нового знания к одному из образов? Как обнаруживать и заполнять пробелы в пространстве знаний?

Рассмотрим краткие ответы на эти вопросы.

3.1. Таксономия знаний. Задача таксономии знаний при наличии меры расстояний между знаниями решается теми же метода - ми, что и задача таксономии данных [3]. В результате, знания объединяются по похожести условий, следствий или продукций в целом. База знаний становится структурированной, что имеет важное значение для ускорения работы экспертной системы: теперь при поиске продукции, похожей на заданную, нет необходимости сравнивать ее со всеми знаниями. Достаточно сравнить с типич - ными представителями всех таксонов и затем провести сравнение только со знаниями самого близкого таксона.

3.2. Выбор информативного подмножества предикатов. Исходным множеством предикатов можно считать список предикатов, которые были упомянуты экспертами хотя бы один раз. Если считать,

что предикат X не зависит от других предикатов, то его информативность можно было бы оценить по корреляции его значений со значениями целевого предиката, если бы мы умели вычислять коэффициент корреляции между предикатами, заданными своими распределениями.

Базовая гипотеза состоит в том, что если предикат X_q информативен, то его малые изменения должны вести к малым изменениям целевого предиката X_s, а большие изменения - к большим. Величина изменения оценивается через расстояние между соответствующими распределениями. Можно взять за основу расстояния R_{qij} и R_{sij} между одноименными предикатами (т.е. X_{qi}, X_{qj} и X_{si}, X_{sj}) из всех пар разных знаний Z_i и Z_j, содержащих предикаты X_q и X_s. По этим двум сериям чисел R_{qij} и R_{sij} можно вычислить модуль коэффициента линейной корреляции, по которому часто судят о зависимости между двумя переменными. Если модуль корреляции высок, значит предикат X_q сильно связан с целевым предикатом X_s и его следует считать важным, информативным.

По-видимому, такой же подход можно использовать и для проверки одновременного влияния двух или большего числа зависящих друг от друга предикатов (X_q, X_t, X_d и т.д.) на целевой предикат X_s. Находятся средние расстояния между одноименными предикатами из левых частей двух знаний и расстояние между целевыми предикатами этих знаний. В результате получается две серии чисел, по которым через модуль их линейной корреляции можно определить зависимость данного набора предикатов на целевой предикат.

Ясно, что большие вычислительные трудности, сопровождающие такого рода NP-полные переборные задачи, в данном случае будут усугубляться сложностью определения расстояний между распределениями. Дополнительная большая сложность может возникнуть, если зависимости носят нелинейный характер, который к

тому же может меняться при разных диапазонах значений предикатов (например, влияние содержания азота в почве на рост растений при низких и при высоких температурах). При этом придется пользоваться методами анализа кусочно-линейных зависимостей [9].

3.3. Распознавание образов в пространстве знаний. Если классификация знаний сделана по значениям целевого предиката, то образы в пространстве знаний будут отличаться друг от друга значениями предикатов из левых частей конъюнкций. Наборы этих предикатов можно заменить эталонной конъюнкцией, составленной из средних значений всех предикатов, встретившихся в описании данного образа.

Здесь возникает нетривиальная задача определения среднего распределения для нескольких распределений. Можно вычислить расстояния между всеми парами распределений и выбрать то распределение, сумма растояний до которого от всех других распределений минимальна. Можно синтезировать "искусственный центр", т.е. построить распределение, сумма расстояний от которого до всех имеющихся минимальна. Для этого случая можно высказать гипотезу (проверенную на ряде примеров, но пока не доказанную) о том, что искусственное "центральное" распределение для предиката X_q можно получить путем механического усреднения плотностей вероятности в каждой градации значений этого предиката для всех знаний данного образа. При распознавании нового знания решение о его принадлежности к j-му образу принимается в том случае, если расстояние между новым знанием и эталоном ("центром") j-го образа минимально.

При небольшом числе знаний образ можно представлять и не прибегая к усредненным эталонам. В этом случае запоминаются все знания из данного образа. Распознаваемое знание сравнивается со всеми знаниями всех образов и относится к тому образу, расстояние до ближайшего представителя которого оказалось наи-

меньшим. Более осторожные (помехоустойчивые) решения принима - ются по методу k ближайших соседей: решение принимается в пользу того образа, чьих представителей оказалось больше среди k самых близких знаний к распознаваемому.

3.4. Заполнение пробелов в знаниях. Список знаний можно записать в форме, близкой к таблице данных типа "объект-свой - ство". Строку i в такой таблице будет занимать знание Z_i, а j-й столбец будет отражать мнения экспертов о значениях предиката X_j. Если информация о значении j-го предиката в строке i от - сутствует, то это значение (X_{ij}) можно попытаться предска - зать с помощью ZET-метода [3], используя закономерные связи между знаниями и между предикатами. Для этого вначале отбира - ется "компетентная" подтаблица размером k на 1. Строка Z_v, v = = 1,2,...,k, включается в число компетентных, если она со - держит информацию о j-м предикате (X_{vj}) и входит в число k наиболее близких (в смысле меры R) к строке Z_i. Аналогично, столбец X_q, q=1,2,...,1, включается в число компетентных, ес - ли известно значение предиката X_{iq} и столбец X_q входит в число 1 наиболее похожих на столбец X_j. Похожесть между столбцами оп - ределяется как среднее расстояние R между принадлежащими од - ной и той же строке (знанию Z_v) парами распределений, одно из которых отражается предикатом X_{vj}, а второе - предикатом X_{vq}.

Затем вычисляются расстояния R_{iv} между строкой Z_i и всеми остальными k строками компетентной подматрицы и синтезируется прогнозное распределение пропущенного предиката X_{lij} как не - которая функция от распределений предикатов X_{vj} с учетом рас - стояний R_{iv}. Это прогнозное распределение должно обеспечивать минимум суммы взвешенных расстояний S от него до всех распре - делений, участвующих в его синтезе:

$$S = \sum_{v=1}^{k} S(X_{lij} - X_{vj}) * (1 - R_{iv})^{\alpha}.$$

Показателем степени α можно регулировать зависимость результата от расстояния R_{iv}: при $\alpha = 0$ все распределения участвуют в вычислениях с равными весами, при больших значениях α будут доминировать распределения, взятые из наиболее близких строк.

Представляется правдоподобным предположение о том, что синтез распределения с указанными свойствами можно сделать путем усреднения значений плотности вероятности в каждой градации значений данного предиката. Если весь диапазон возможных значений предиката X_{vj} разделен на m градаций и вероятность того, что предикат принимает значение d-й градации равна P_{vjd}, то усредненное по всем строкам значение плотности в этой градации будет равно:

$$P_{jd} = \{ \sum_{v=1}^{k} P_{vjd} * (1 - R_{iv})^{\alpha} \} / \sum_{v=1}^{k} (1 - R_{iv})^{\alpha}.$$

Еще один прогноз X_{2ij} можно получить, используя зависимости между j-м и всеми l другими столбцами (предикатами) компетентной подтаблицы. Здесь суммировать с весами R_{jq} нужно распределения всех предикатов X_{iq} строки i. В качестве окончательного прогноза распределения пропущенного предиката X_{ij} можно принять усредненное по градациям значение двух полученных прогнозов X_{1ij} и X_{2ij}.

Для оценки величины ожидаемой ошибки можно, как и в алгоритме ZET, применить метод контрольного прогнозирования известных значений предикатов в компетентной подтаблице.

З а к л ю ч е н и е

Практическое использование описанных выше методов анализа знаний окажется возможным, если будет найден корректный и технически несложный способ синтеза "средних" распределений. Проблема будет решена, если удастся доказать правильность описанной

в работе гипотезы о возможности механического усреднения плотностей для каждой отдельной градации, т.е. если удастся доказать следующую теорему

ТЕОРЕМА. *Если расстояние* R *между* K *распределениями измерять по формуле* (*1*), *диапазон возможных значений переменной ограничен и разделен на* m *градаций, то определяя плотность в каждой градации как среднеарифметическую величину от плотностей в данной градации всех* T *распределений, мы получим распределение* $X_{ср}$, *сумма расстояний от которого до всех этих распределений будет минимальной.*

Если будет показано, что гипотеза не верна, то потребуется дальнейшая работа по поиску метода синтеза "средних" распределений.

Литература

1. ЗАГОРУЙКО Н.Г. Классификация задач анализа прогнозирования на таблицах "объект-свойство" //Машинные методы обнаружения закономерностей. - Новосибирск, 1981. - Вып.88: Вычислительные системы. - С. 3-8.

2. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. - М.: Советское Радио, 1972. - 206 с.

3. ЗАГОРУЙКО Н.Г., ЁЛКИНА В.Н., ЕМЕЛЬЯНОВ С.В., ЛБОВ Г.С. Пакет прикладных программ ОТЭКС.-М.: Финансы и статистика,1986. - 160 с.

4. СУПЕС П., ЗИНЕС Дж. Основы теории измерений //Психологические измерения. - М.: Мир, 1967. - С. 117-132.

5. ЗАГОРУЙКО Н.Г., БУШУЕВ М.В. Меры расстояния в пространстве знаний //Анализ данных в экспертных системах. - Новосибирск, 1986. - Вып. 117: Вычислительные системы. - С. 24-35.

6. ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г. Блок анализа данных в экспертной системе ЭКСНА //Экспертные системы и анализ данных. - Новосибирск, 1991. - Вып. 144: Вычислительные системы. -С. 54-175.

7. БУШУЕВ М.В., ЁЛКИНА В.Н., ЗАГОРУЙКО Н.Г., ШЕМЯКИНА Е.Н. Блок анализа знаний в инструментальной экспертной системе ЭКСНА //Методы и системы искусственного интеллекта. - Новоси - бирск, 1992. - Вып. 145: Вычислительные системы. - С.29-79.

8. ЗАГОРУЙКО Н.Г. Согласование разнотипных шкал //Анализ разнотипных данных. - Новосибирск, 1993. - Вып. 99: Вычислительные системы. - С. 3-14.

9. РОЗИН Б.Б., КОТЮКОВ В.И., ЯГОЛЬНИЦЕР М.А. Экономико статистические модели с переменной структурой. - Новосибирск: Наука, 1984. - 241 с.

Поступила в ред.-изд.отд.

23 мая 1994 года