Курс лекций «Введение в ИИ.». Часть II. Распознавание образов
Лекция 12. Элементы теории измерений. Таксономия

О.Г. Чанышев

Содержание

1 Элементы теории измерений

Если вы ведете протокол измерений «объект-свойство» для собственного пользования, то вы сободны в выборе языка ведения протокола. Только сами помните, что измеряли и как кодировали результаты измерений. Если же протокол будет использоваться другими людьми, а тем более, если он предназначен для использования компьютерными программами, требуется обеспечить однозначное понимание смысла любым пользователем.

Из множества возможных способов отображения наблюдаемого мира получили распространение всего несколько, ставшие общепринятыми. Их изучением занимается *теория измерений*.

В настоящей лекции представляются основные сведения из этой теории, необходимые для описания методов анализа данных.

1.1 Типы измерительных шкал

В процессе измерения участвуют два объекта: измерительный прибор и измеряемый объект. Теория измерений оперирует понятиями *эмпирическая система с отношениями* и *символьная система с отношениями*

$E = \{A, R\}$ - эмпирическая система, где A - множество измеряемых объектов, R - отношения между ними.

$N = \{M, P\}$ - символьная система, где M - множество сиволов, P - конечный набор отношений на этих символах.

Договоренность использовать фиксированное отображение системы E на N означает выбор некоторого определенного правила отображения g. Тройка $\langle E, N, g\rangle$ называется *шкалой*.

Мы можем договриться о некотором другом способе отображения w, и тогда будем иметь дело с другой шкалой $\langle E, N, w\rangle$. Например, в g записываем вес в килограммах, а в w - в тоннах. Цифровая запись оказывается различной при одинаковом эмпирическом содержании. Это означает, что выбраны не любые способы отображения, а только такие, которые связаны взаимно однозначными преобразованиями. Т.е. имеется такое преобразование f, что

$$g = f(w),\ w = f'(g)$$

Преобразование f объединяет шкалы в группу, которая называется *типом шкалы*.

В практике научных исследования получили распространение шкалы всего нескольких типов.

1. *АБСОЛЮТНАЯ ШКАЛА*. Допустимое преобразование для шкал данного типа представляет из себя тождество. Пример: 6=VI.

2. *ШКАЛА ОТНОШЕНИЙ*. Между разными протоколами, фиксирующими один и тот же эмпирический факт на разных языках должно выполняться отношение $y = \alpha \times x$.

   Один и тот же смысл имеют протоколы: 16 кг, 16 000 г, 0.016 т, 1 пуд.

   Этот тип шкалы удобен для записи величин, имеющих единицу измерения - длин, весов, скоростей и т.п.

3. *ШКАЛА ИНТЕРВАЛОВ*. Здесь между протоколами y x допустимы линейные преобразования $y = \alpha \times x + b$. Это означает, что в разных протоколах может использоваться разный масштаб единиц (α) и разные начала отсчета (b). Пример: шкалы для измерения температуры (Цельсия,Кельвина, Фаренгейта).

4. *ШКАЛА ПОРЯДКА*. Допустимыми в данных шкалах являются все монотонные преобразования, т.е. такие, которые не нарушают порядок следования значений измеряемых величин. Такие протоколы появляются, например, при сравнении тел по твердости. Записи «1,2,3» и «5.3, 12.5, 109.2» содержат одинаковую информацию о том, что первое тело является самым твердым, второе - менее, а третье - еще менее твердым.

   Разновидностью шкал порядка являются шкалы рангов, идущие от 1 вверх по возрастанию (например, ранги слов в тексте с совпадающими частотами). Разновидностью шкал порядка являются шкалы баллов. При этом используются целы числа в ограниченном диапазоне значений: от 1 до 5 или от 1 до 12 (оценки успеваемости в школе), от 0 до 6 или 10 (в спорте).

   Протоколы шкал порядков содержат информацию только о трех эмпиричеких отношениях «больше», «меньше» и «равно».

5. *ШКАЛА НАИМЕНОВАНИЙ.* В этих шкалах фиксируются только два отношения «равно» и «не равно». Следовательно, допустимы любые преобразования, лишь бы разные объекты имели разные обозначения (имена). Примеры: национальность, названия населенных пунктов и т.п. (Если два разных населенных пункта имеют одинаковые названия, то добавляются уточнители, которые уже вместе с названием являются уникальным наименованием).

1.2 Сравнительная информативность шкал

Представляет интерес вопрос об относительной информативности шкал. С позиций порядковой шкалы ответ ясен: информативность шкал убывает в порядке их перечисления в данном тексте.

Пример.

Пусть в протоколе, записанном в абсолютной шкале содержится информация о том, что множество A содержит 30 элементов, а множество B - 10.

На языке шкалы отношений будет зафиксировано, что множество A содержит в три раза больше элементов, чем множество B.

Шкала порядка зафиксирует, что множество A содержит больше элементов, чем множество B.

Шкала наименований может зафиксировать лишь тот факт, что множества A и B содержат различное число элементов.

Таким образом, информации, содержащейся в абсолютной шкале, достаточно для ее однозначного отображения на более слабую шкалу. Обратное не верно. Из факта $A \neq B$ нельзя сказать, какое из множеств больше, во сколько раз одно множество больше другого и, тем более, по сколько элементов содержится в A и B.

Информацию, записанную в шкалах первых трех типов можно подвергать математическим преобразованиям. Поэтому их часто называют *сильными*, *количественными* или *математическими* шкалами.

Шкалы порядка и наименований называют *слабыми* и *качественными*.

Нельзя рекомендовать пользоваться шкалами только первых трех типов. Дело в том, что приборы для измерения сильных свойств более дорогие, а знания во многих областях (например, в гуманитарных) имеют качественный характер. Или, иными словами, нам недостает знаний для измерения свойств объектов в этих областях в сильных шкалах.

Представление о том, как много информации мы теряем, переходя от сильных шкал к более слабым, можно получить, оценив количество неизоморфных (разных) протоколов в разных шкалах.

Будем считать, что измерительный прибор может принимать одно из m состояний. Пусть этим прибором измеряется фиксированное свойство у $n > 1$ объектов. Тогда протоколы «2, 6», «3, 9» в абсолютной шкале неизоморфны, а в шкале отношений, порядка и наименований - изоморфны.

По этой методике было проведено сравнение шкал трех типов: абсолютной, порядковой и наименований. Сравнение проводилось в шкале отношений: количества неизоморфных протоколов для шкалы порядка (S_o) и шкалы наименований S_n соотносилось с количеством неизоморфных протоколов в абсолютной шкале (S_a.) Выяснилось, что для фиксированного значения числа градаций m с ростом количесва измеряемых объектов n различия в информативности шкал уменьшаются. Однако, отношение $\frac{S_n}{S_a}$ остается малым и меняется слабо. Отношение же $\frac{S_o}{S_a}$ растет быстро и при $n > 5 \times m$ достигает величины 0.9. Т.е. информативность шкалы порядка

при экспериментах с большим числом объектов приближается к информативности абсолютной шкалы.

Так что в ряде случаев при использовании более простых приборов и процедур можно получить почти столько же информации, сколько с помощью сложных и дорогих.

Пример.

28 экспертов должны были оценить некоторое неформализованное свойство 10 объектов в шкале порядка. Каэждый эксперт упорядочивал объекты по своему усмотрению и приписывал им целочисленные порядковые значения в диапазоне от 1 до 10. Затем им было предложено оценить свойство тех же объектов в шкале отношений (в процентах к самому лучшему). Всеми эксператми эта задача оценвалась как существенно более трудная. После завершения работы были определены для каждого объекта средние значения их порядковых мест и средние значения процентных оценок. Оказалось, что коэффициент линейной корреляции между этими средними оценками составляет 0.93!

Отсюда можно сделать полезный вывод для практики группового экспертного оценивания:*не нужно заставлять экспертов давать ответы в сильных шкалах.* При количестве экспертов около 30 достаточно ограничиться оценками в шкале порядка. И лишь для двух объектов, получивших самый высокий и самый низкий порядковый балл, сделать оценку в сильной шкале. Этих калибровачных величин будет достаточно для перехода от средних значений в шкале порядка к средним значениям в шкале отношений.

2 Таксономия. Алгоритмы класса FOREL

Человек, группируя объекты, руководствуется некоторым критерием (F). Следовательно, гипотеза более сильная, чем H, должна быть сформулирована с учетом этого критерия. Тестовый алгоритм должен допускать только такую группировку, которая удовлетворяет критерию F. Иными словами, следует конкретизировать понятие сходства, «схожести» объектов.

Считаем, что признаки объектов заданы в сильных шкалах и мы можем работать в метрических пространствах. В частности, можем в евклидовом многомерном пространстве признаков ввести расстояние между точками.

Пусть

$C^j = (x_1^j, x_2^j, .., x_i^j, .., x_n^j$ - координаты центра тяжести j-го таксона,

$\rho_i^j(C^j, a_i)$ - расстояние между центром тяжести и произвольной точкой a_i,

$\rho^j = \sum_{i=1}^{m_i} \rho_i^j(C^j, a_i)$ - сумма таких расстояний j-го таксона.

$F = \sum_j^k \rho^j(C^j, a_i)$.

Смысл критерия похожести на центр состоит в том, чтобы найти такое разбиение m объектов на k таксонов, при котором F минимальна.

2.1 Алгоритм Forel

1. Признаки объектов нормируются так, чтобы их значения находились между нулем и единицей.

2. Строим гиперсферу минимального радиуса R_0, охватывающую все m точек.

3. $R_0' = 0.9R_0$

4. Помещаем центр сферы в любую из внутренних точек (расстояние до которых меньше радиуса) и вычисляем их центр тяжести.

5. Переносим центр сферы в центр тяжести и снова находим внутренние точки.

Таким образом, центр сферы перемещается в область локального сгущения точек.

Когда сфера остановится, внутренние точки объявляем принадлежащими таксону № 1 и исключаем их из рассмотрения.

Повторяем процедуру с оставшимися точками, и так до тех пор, пока все точки не будут распределены по таксонам.

Если начальную точку на шаге 4 менять случайным образом, может получиться несколько вариантов таксономии, из которых выбирается тот, на котором достигается $min(F)$.

2.2 Forel-2

Эта модификация исходного алгоритма используется в случае, когда нужно получить в точности t-таксонов. Радиус сферы по мере надобности уменьшается или увеличивается на величину R, которая от итерации к итерации уменьшается, например, вдвое.

Наилучшему варианту таксономии отвечает $min(F)$ при числе таксонов равном t.

2.3 SKAT

В результате работы алгоритма FOREL могут образоваться неустойчивые таксоны, случайные сгустки, тяготеющие к одному из полученных таксонов. Это может произойти в результате недостаточной «корректности» исходных данных.

Один из эвристических приемов для учета неустойчивых таксонов реализован в алгоритме SKAT.

На вход программы поступает исходное множество объектов m и результат таксономии по алгоритму FOREL S. Процесс повторяется с тем же радиусом сфер, но на всех m точках множества и с начальными точками, совпадающими с центрами таксонов.

2.4 Алгоритм KOLLAPS

Алгоритм KOLLAPS применяется в задачах выделения локальных сгустков точек из фона. Например, выделения ярких созвездий на фоне неба.

На первом этапе решения задачи определяются центры сгустков, а на втором - проверяется, действительно ли эти точки являются центрами устойчивых таксонов.

1. Прежде всего задается некоторое значение порога плотности - количества точек m_j в таксоне (d). Если $m_j > d$, то запоминается центр таксона, а его точки из дальнейшего рассмотрения исключаются. В противном случае центр не запоминается, но точки таксона из рассмотрения исключаются.

2. Затем центр переносится в любую из оставшихся точек и процесс продолжается до исчерпания всех точек.

3. Восстанавливаем все множество точек, выбираем таксон с максимальным значением d, помещаем сферу в его центр.

4. Начинаем сжимать сферу.

5. На каждом шаге сжатия определяем число внутренних точек. Если начальный радиус был слишком велик, то изменение d будет медленным. По мере вхождения в более плотные области, темп изменения будет увеличиваться, что и служит сигналом к остановке сжатия. Фиксируется число внутренних точек таксона m'_j.

6. Процедура повторяется для всех запомненных центров таксонов.

7. Выбирается k таксонов с наибольшим количеством точек.

2.5 Алгоритм BIGFOR

Что делать, если массив из m точек очень велик и не помещается в оперативной памяти? При этом затраты на «поточечное» чтение координат из внешнего носителя будут неприемлемо велики.

1. Разбить исходный массив на $t = \frac{m}{V}$ подмассивов, V - число точек в подмассиве. m С помощью FOREL-2 разделить каждый из подмассивов на k' таксонов. Описание каждого j-го таксона содержит координаты его центра и количество внутренних точек m'_j.

   В итоге получим $q = tk'$ точек-центров таксонов.

2. Вновь используем FOREL-2 для разбиения этих точек на k-таксонов. Только при расчете центров тяжести учитывается вес («массу») m'_j.

3. Перераспределяем точки между k таксонами.

2.6 Иерархическая таксономия

Для иерархической таксономии по методу «снизу-вверх» используется алгоритм BIGFOR, только с исключенной процедурой перераспределения точек.

На первом шаге радиус устанавливается малым, дающим таксоны нижнего уровня. На последующих шагах таксоны укрупняются, образуя вышележащие урони иерархии. Процесс прекращается, когда в итоговый таксон войдут все точки исходного множества.

Для кластеризации «сверху-вниз» используется FOREL с последовательным уменьшением радиуса сферы.

1. Определяем минимальный радиус гиперсферы R, включающей все точки m. Эти точки составляют таксон верхнего уровня.

2. Уменьшая от шага к шагу радиус, определяем таксоны i-х уровней.

3. Процесс завершается при числе таксонов нижнего уровня равного m - по точке в таксоне.

2.7 Динамическая таксономия

При зависимости мощности кластеризуемого множества от времени результат кластеризации может меняться с появлением или исчезновением точек.

Для таксономии объектов, возникающих по одному или небольшими группами применяется алгоритм DINA.

Задается некоторый радиус R. Первая появившаяся точка или группа точек объявляется центром первого таксона. При появлении новой точки производится проверка, попадает ли точка внутрь гиперсферы. В зависимости от результата точка либо включается в состав таксона, а центр гиперсферы смещается в центр тяжести внутренних точек, либо новая точка объявляется центром нового таксона. Далее процесс очевиден.

Можно следить за тем, чтобы таксоны не «переполнялись» - содержали по возможности одинаковое количество точек. При переполнении таксон можно разбить на два с одинаковым числом точек.

Переход от описания исходных объектов к описанию таксонов эквивалентен переходу от данных к знаниям. Иерархическая таксономия отображает структуру нашего знания о изучаемом явлении. Можно строить иерархии понятий («растущие пирамидальные сети») в процессе накопления новых фактов. Могут возникать таксоны с чрезмерно большим количеством объектов и тогда их следует «таксономировать», что эквивалентно детализации знаний.

2.8 Таксономия с суперцелью. Алгоритм ROST

Рассмотренные выше алгоритмы таксономии - это универсальные алгоритмы, так сказать, «на все случаи жизни». Однако, такой подход во многих конкретных случаях может оказаться неоптимальным.

Например, в задачах распознавания устной речи, обычно пользуются не отдельными фонемами, а их группами (звукотипами). При этом неприемлемо, если в одну группу попадут очень похожие по своим спектральным характеристикам звуки «т», «к», «п» («ток», «кот», «кто»). Таксоны на уровне звукотипов должны строиться с учетом «суперцели»: помимо того, что таксоны должны объединять похожие элементы,количество таксонов должно быть минимальным, но достаточным для принятия решений на более высоких уровнях.

Таким образом, мы четко сформулировали цель таксономии и термины типа «самообучение», «обучение без учителя» более неприемлемы.

Таксономия с учетом суперцели может быть получена алгоритмом ROST.

Алгоритм ROST - это вариант иерархической таксономии методом «снизу вверх». Вначале мы применяем FOREL с малым радиусом гиперсферы. Затем радиус увеличиваем и после каждого шага делаем проверку на соответствие суперцели: не возникают ли ошибки из-за укрупнения таксонов (например, звукотипов). Если нет, то процесс продолжается, в обратном случае, то точки таксона, приводящего к ошибкам подвергаются повторной таксономии с меньшим радиусом гиперсферы и из дальнейшего рассмотрения исключаются. Укрупнение продолжается до тех пор, пока из рассмотрения не будут исключены все точки.

3 Метод сравнения алгоритмов таксономии

Главное, что интересует пользователя - качество полученных решений. Для того, чтобы сформулировать критерий качества, вспомним о том, что таксономия производится не только и не столько для компактного преобразования множества m объектов в k таксонов. В дальнейшем эти таксоны или их типичные представители используются для краткого описания имеющихся объектов и, что более важно, для распознавания новых объектов генеральной совокупности. Каждый новый объект относится к наиболее «близкому» таксону (образу).

Пусть M - мощность генеральной совокупности, $m < M$ - мощность некоторой выборки из генеральной совокупности. На множестве выборки произведена таксономия некоторым алгоритмом F. Если теперь предъявлять программе классификации остальные $M - m$ объектов и присоединять их к уже полученным таксонам, получим вариант таксономии генеральной совокупности S'. Если тем же алгоритмом F сделать таксономию объектов всей генеральной совокупности, то получим вариант S. Таксономии S и S' назовем соответственно *базовой* и *выборочной*.

Если базовая и выборочная таксономии совпадают, то алгоритм F удачно угадал структуру генеральной совокупности по случайной выборке. Способность по малым выборкам правильно угадывать структурные закономерности генеральной совокупности и есть основная характеристика качества (Q) таксономии.

Пусть p и q - объекты. Введем понятие «таксономического расстояния»:

$$\rho(p,q) = \begin{cases} 1, & \text{Если p и q принадлежат разным таксонам} \\ 0, & \text{Если p и q принадлежат одному таксону} \end{cases}$$

Пусть в квадратной матрице размерностью $M \times M$ столбцы и строки соответствуют объектам генеральной совокупности,а на пересечении p-ой строки с q-м столбцом находится значение $\rho(p,q)$. Матрица симметрична, а диагональные элементы равны 0. Различие $R(S,S')$между таксономиями можно получить, просуммировав число элементов с несовпадающими значениями у двух матриц, представляющих результаты S, S' таксономий. Разделив полученную сумму на максимально возможное число несовпадений, которое равно $(M^2 - M)$, получим расстояние Хэмминга между матрицами, нормированное в диапазоне от 0 до 1:

$$R(S,S') = \sum_{p,q=1}^{M} \frac{r(p,q) - r'(p,q)}{M^2 - M}$$

Чем меньше значение $R(S,S')$, тем лучше алгоритм F угадал структуру генеральной совокупности.

Для того, чтобы практически реализовать проверку качества, можно написать программу, которая генерировала бы данные с заданным законом распределения свойств в пространстве признаков, мощности генеральных совокупностей и случайных выборок. Такая программа («полигон Таксон») разработана и проверка различных алгоритмов дала следующие результаты.

Во-первых, лучшим алгоритмом оказался KRAB, на втором месте был SKAT и на третьем - FOREL. Но FOREL дает быстрые и простые решения.

Во-вторых, выяснилось, что качество таксономии не зависит от размерности признакового пространства. Можно сказать, что по этому свойству машина значительно превосходит человека, который решает успешно задачи таксономии, лишь когда непосредственно видит разделяемое множество. Иначе он переходит к делению по каждому признаку в отдельности. (На мой

взгляд, это связано с ограниченностью объема кратковременной памяти; ничего удивительного в полученном результате нет, поскольку для этого (быстродействующая мельница чисел) и создавались ЭВМ - О.Ч.)

Веса признаков В реальных задачах, когда мы не можем точно определить важность той или иной признаковой координаты, другими словами, определить размерность признакового пространства, следует сделать предположение (выдвинуть гипотезу) о важности того или иного признака и формально учесть это, присваивая признакам некоторые значения весов. Тогда евклидово расстояние между объектами в n-мерном признаковом пространстве:

$$\rho(p,q) = \sum_{i=1}^{n} \sqrt{\gamma_i (x_i^p - x_i^q)}$$

Следующая лекция

Лекция 13. Алгоритмы распознавания образов