МАШИННЫЕ МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
1981 год Выпуск 88

УДК 519.95:681.3.06

СКОРОСТНЫЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Л.К.Выханду

Успехи в деле обработки данных в большой степени зависят и от степени полезности той информации, которую извлекают исследователь на чередующихся этапах осмысления полученных результатов в процессе обработки данных. Обычно результаты обработки данных представлены набором чисел, воспринять которые в короткий промежуток времени затруднительно, и исследователь, как правило, начинает систематизировать представленные в наборе числа, полагаясь при этом полностью на собственную интуицию.

Задача систематизации объектов в литературе известна как задача ординации объектов, т.е. упорядочения их в каком-то естественном порядке.

В данной работе предложено новое семейство быстрых методов ординации, строящихся на понятии эмпирических частот матрицы данных объект-признак и на понятии взаимовлияния объектов, объекта и признака или признаков.

I. Частотные методы упорядочения

Новое семейство методов ординации образуется, если от матрицы данных A перейти к частотным таблицам данных при помощи так называемого частотного преобразования.

Формально такой переход можно осуществить следующим образом. Для матрицы данных A , состоящей из значений признаков объектов, измеряемых по номинальной или ординальной шкале, для каждого признака составляются эмпирические гистограммы - эмпирические функции распределения

$$f_j = (f_{1j}, f_{2j}, \ldots, f_{l_j j}), \quad \sum_{R=1}^{l_j} f_{hj} = N ,$$

где l_j - количество классов для значений j-го признака. Если признак j интервальный, то промежуток изменения признака делится на l_j равных участков. После необходимых преобразований и вычисления всех гистограмм $f_j (j = 1, \ldots, M)$, заменив каждое значение $a_{ij} = h$ j-го признака в i-м объекте матрицы данных на частоту f_{hj}, получим новую матрицу Z, которую назовем частотной. Переход от матрицы данных A к частотной матрице обозначим через $A \to Z$, а элементы Z - через z_{ij}. Будем говорить, что $A \to Z$ есть частотное преобразование матрицы A.

Смысл данного частотного преобразования состоит в следующем. Во-первых, сумма частот каждого объекта дает какую-то оценку его конформности (обычности) по отношению ко всей группе объектов; в группе легко обнаружить вероятного максимального (минимального) конформиста. Для максимального конформиста, который согласуется с группой по каждому признаку, надо просто найти максимальную частоту каждого признака (так называемую моду), сложив все эти моды, чтобы получить число U. Во-вторых, сумма L минимальных, отличных от нуля частот в каждом признаке указывает предел, который характеризует минимальные для изучаемой группы значения признаков. Отрезок [L,U] назовем шкалой конформизма [1].

Сумма частот каждого объекта представляет собой градацию на шкале конформизма - на отрезке [L,U] . Объекты с наибольшими значениями на этой шкале являются наиболее типичными для группы объектов матрицы A; объекты с наименьшими значениями - наименее типичные объекты группы.

Таким образом, из матрицы данных A на основе частотного преобразования получается новая матрица Z тех же размеров, которую уже можно обрабатывать как матрицу, заданную в интервальной шкале. Шкала конформизма открывает многочисленные возможности упорядочения множества объектов по значениям этой шкалы.

Одновременно упорядочение можно осуществить и на множестве признаков. Для этого на основе гистограмм каждого признака вычисляются суммы $H_j = \sum_{h=1}^{k} f_{hj}^2$, которые в литературе [2] обычно служат измерителями варьируемости признаков по мере возрастания (убывания) сумм H_j. В результате происходит двустороннее упорядоче-

ние матрицы данных в системе объектов и признаков. Исследователь получает в свое распоряжение такое представление матрицы данных A , которое более отчетливо выделяет особенности материала на случай, если какой-либо объект "резко" отличается от других.

В заключение данного пункта установим требуемое количество операций для упорядочения объектов в шкале конформизма. Количество вычислений для получения градаций объектов на шкале конформизма (т.е. число операций для получения гистограмм) равно N×M, и столько же операций необходимо для вычисления градаций. Следовательно, полное упорядочение объектов на шкале конформизма относится к скоростным методам, требующих O(N M) операций.

2. Методы упорядочения ортогональными преобразованиями

Определение значения объекта на шкале конформизма как суммы частот в заданной системе объектов, кроме удовлетворительной содержательной трактовки, представляет еще возможность эффективно использовать развитую в [3] методику скоростных преобразований.

Допустим, что осуществлен переход $A \to Z$ и матрица данных A представлена в частотном виде Z . Легко показать, что после преобразования ZH , где H - матрица преобразования Адамара (или Хаара), первый столбец матрицы $\overline{Z} = ZH$, будучи суммой строк Z, совпадает с градациями на шкале конформизма системы объектов A . Следовательно, на основе результатов [3] можно утверждать, что основная доля варьируемости объектов в системе A заключена в первом столбце $\overline{Z} = ZH$. Ниже приводятся формальные определения, обобщающие определение шкалы конформизма.

Предположим, что собрана информация о множестве из N объектов.

Пусть многомерное признаковое пространство, в котором заданы эти объекты, описывается M-дискретными случайными переменными $(x_1, x_2, \ldots, x_M)$. Вектор-строка $a_i = (a_{i1}, a_{i2}, \ldots, a_{iM})$ матрицы A состоит из множества значений случайных переменных, соответствующих i-му объекту, вектор-столбец $a^j = (a_{1j}, a_{2j}, \ldots, a_{Nj})$ представляет множество значений j-й случайной переменной x_j.

Допустим, что каждая переменная x_j имеет алфавит L_j возможных значений $\alpha_{jh} \in L_j$, $h = 1, 2, \ldots, l_j$; l_j - количество букв в алфавите L_j. Эмпирическая частота события $x_j = \alpha_{jh}$ в системе объектов A выражается в виде $f_{hj} = \sum_{i=1}^{N} \delta_{ijh}$,

где

$$\delta_{ijh} = \begin{cases} 1, & \text{если} \quad a_{ij} = \alpha_{jh}, \\ 0, & \text{если} \quad a_{ij} \neq \alpha_{jh}. \end{cases}$$

Эмпирические частоты f_{hj} дают разнообразные оценки вероятности P_{hj} осуществления события $x_j = \alpha_{jh}$. Например, в социологии обычно используют оценки $P_{hj} = f_{hj}/N$. В работе американских статистиков [5] предложены так называемые квазибайесовские оценки, которые стабильнее обычных и у которых функция квадратного риска меньше.

В качестве эффективно вычисляемого измерителя роли шкалы конформизма для системы A можно, например, предложить отношение $\tilde{f}_1 / \sum_{j=1}^{M} \tilde{f}_j$, где $\tilde{f}_j = \sum_{i=1}^{N} \tilde{z}_{ij}^2$ и $\tilde{z}_{ij}$ есть элемент матрицы $\tilde{Z} = ZH$. Отношение $\tilde{z}_1 / \sum_{j=1}^{M} \tilde{z}_j$ показывает, какую долю варьируемости частот матрицы Z описывает шкала конформизма.

Вообще говоря, матрица $\tilde{Z}=ZH$ удобна для обобщения шкалы конформизма с двух точек зрения. Во-первых, методом частичной ортогонализации [3] можно достичь бо́льшей варьируемости частот в проекции на главное направление матрицы Z (первый столбец $\tilde{Z}$). В этом смысле частичную ортогонализацию Z можно понимать как уточнение шкалы конформизма. Во-вторых, осуществляя переход на двумерное (визуальное) представление частотной информации об объектах, можно выбрать два столбца $\tilde{Z}$ с наибольшими суммами квадратов f_j; выбранная пара столбцов Z задает координаты объектов на плоскости. Эти координаты и определят визуальную картину, которая естественным образом раздвигает шкалу конформизма одного измерения на случай двух измерений.

Отметим, что осуществление стандартной схемы вычислений матрицы расстояний и двух главных собственных векторов для получения аналогичной визуальной картины расположения объектов требует в N раз больше затрат, чем схема, приведенная выше.

3. Методы упорядочения в нелинейном случае

Рассмотрим нелинейные обобщения шкалы конформизма. Как было показано выше, для уточнения шкалы конформизма могут применяться ортогональные H-преобразования, осуществляемые переходом от матрицы данных A к ее частному виду Z.

Определим меру вариации каждого объекта (не проводя H-преобразований) как сумму $S_i = \sum_{j=1}^{M} Z_{ij}^2$. Чем больше S_i, тем конформнее объект.

Далее, исходя из меры вариации S_i каждого объекта, можно найти и меру вариации всей их системы как число $S = \sum_{i=1}^{N} S_i$. С одной стороны, ввиду аддитивности меры S равенство $S = \sum_{j=1}^{M} \sum_{i=1}^{N} z_{ij}^2$ позволяет оценить долю варьируемости частот каждого признака по системе объектов A. С другой стороны, сам объект оказывает влияние на варьируемость частот группы объектов, к которой этот объект не принадлежит.

Определим влияние объекта i на совокупность объектов как изменение, вносимое в сумму квадратов S, если из системы исключить этот объект, полагая все его признаки неопределенными. При этом изменения, вносимые в S самим исключаемым объектом, не учитываются.

Для признака j по значению $h = a_{ij}$ определяется соответствующая эмпирическая частота f_{hj}. Поскольку $Z_{ij} = f_{hj}$ и каждая такая частота f_{hj} в множестве объектов $A\backslash i$ имеет еще $f_{hj} - 1$ объектов с равными h значениями признака j, то сумма квадратов частот из-за неопределенности значения $a_{ij} = h$ уменьшается на

$$\pi_{ij} = (Z_{ij}-1)(Z_{ij}^2-(Z_{ij}-1)^2) = 2Z_{ij}^2 - 3Z_{ij} + 1.$$

Заметим, что в π_{ij} не учитывается уменьшение величины S, вызываемое исключением объекта i. В указанном смысле π_{ij} представляет собой чистое влияние на совокупность объектов $A\backslash i$. Суммируя по всем признакам, определим числом $\pi(i) = \sum_{j=1}^{M} \pi_{ij}$ меру влияния объекта на систему объектов A. Совокупность чисел $\pi(i)$ (i = 1,...,N) образует шкалу влияния объектов.

Теперь выявим зависимость между шкалой конформизма и шкалой влияния. Раскроем выражение

$$\pi(i) = \sum_{j=1}^{M} \pi_{ij} = \sum_{j=1}^{M} (2Z_{ij}^2 - 3Z_{ij} + 1) = 2\sum_{j=1}^{M} Z_{ij}^2 - 3\sum_{j=1}^{M} Z_{ij} + M = 2S_i - 3K_i + M.$$

Таким образом, мера влияния объекта состоит из линейной комбинации варьируемости частот и конформизма. Ввиду квадратичности S_i как функции частот шкала влияния "усиливает" более конформные объекты, ослабляя неконформные. Предложенный метод шкалирования объектов в N раз (по количеству операций) эффективнее традиционного метода главных компонент.

Следует также отметить, что аддитивность шкалы влияния позволяет вновь воспользоваться техникой ортогонализации частот. Отправной точкой служит частотная матрица Z, которая заменяется на матрицу $\Pi = (\pi_{ij})$. Матрица Π подвергается H-преобразованию с последующей частичной ортогонализацией.

Рассмотренным здесь методом шкалирования объектов можно воспользоваться и для шкалирования признаков, следует лишь поменять признаки и объекты ролями или, другими словами, применить метод шкалирования к матрице A^T. Однако следует учесть, что непосредственно к матрице данных A^T этот метод неприменим, поскольку гистограммы f_{hj} составляются для признака j и не имеют смысла (ввиду номинальности шкал измерения) для объектов. В следующем разделе к матрице данных A предлагается так называемое унифицирующее преобразование, открывающее пути использования шкалирования также и признаков по предложенной выше методике.

4. Унификация матриц номинальных данных

Выше для анализа матрицы номинальных данных A было предложено частотное преобразование $A \to Z$. Матрица $Z = N \times M$ - матрица с элементами z_{ij}, каждый из которых равен числу объектов (частоте) f_{hj} со значением признака j, равным a_{ij} - элементу матрицы данных A. Одновременно с преобразованием $A \to Z$ рассматривалось ортогональное преобразование $\tilde{Z} = ZH$, которое обеспечило как получение одномерной шкалы конформизма объектов, так и удобную двумерную визуализацию объектов при помощи двух столбцов $\tilde{Z}$ с наибольшими суммами квадратов. Мы сознательно оставили в стороне вопрос о "возмущающем воздействии" на упорядочение и визуализацию различного количества градаций l_j признака j. Поэтому ниже для снижения воздействий различных l_j $(j=1,2,\ldots,M)$ предлагается более общее частотное преобразование, при котором каждое значение a_{ij} заменяется величиной $S_j \cdot Z_{ij}^q$.

Осуществим выбор коэффициентов S_j, $j = 1,2,\ldots, M$, и параметра q так, чтобы при равномерном распределении значений всех

признаков суммы квадратов величин $S_j^2 \cdot Z_{ij}^{2q}$ в столбцах матрицы Z были равными. Этим достигается стандартизация варьируемости признаков, часто используемая в интервальных шкалах данных.

В интервальных шкалах переменная x_j преобразуется в переменную $\bar{x}_j$ так, чтобы $\sum_{i=1}^{N} \bar{x}_{ij} = 0$ и $\sum_{i=1}^{N} \bar{x}_{ij}^2 = 1$. Подобный прием исключает необходимость рассмотрения единиц измерения, и тем самым все признаки оказываются непосредственно сравнимыми. В номинальных шкалах арифметические действия недопустимы, и предлагаемый вариант нормировки исключает здесь появление нежелательных искажений, вызываемых различным количеством градаций признаков, что равносильно привнесению соизмеримости в различные градации номинальной шкалы.

Из условия нормировки явствует, что коэффициент S - это функции количества градаций l. Выберем S в форме функции l^t, где t - переменная величина. Тогда при равномерном распределении значений признака сумма квадратов преобразованных значений z_{ij} признака j матрицы Z равна числу

$$N\left(l^t\left(\frac{N}{l}\right)^q\right)^2 = N^{2q+1} \cdot l^{2(t-q)} .$$

Эта сумма не зависит от l, если $t = q$, т.е. преобразованием $a_{ij} \to l_j^q \cdot z_{ij}^q$ при равномерном распределении обеспечивается равенство сумм квадратов для всех преобразованных признаков независимо от количества градаций. Сумма квадратов равна N^{2q+1}, а деление значений признака на эту величину как раз и дает желаемую нормировку вариации для каждого признака. Наиболее элементарно интерпретируемые варианты нормировки получаются при $q = 1,2$.

В случае преобразования $A \to Z_1 = (l_j z_{ij})$ матрица Z_1 называется матрицей нормированных частот, а в случае преобразования $A \to Z_2 = (l_j^2 z_{ij}^2)$ матрица Z_2 называется матрицей нормированных вариаций.

Теперь для анализа структуры матриц Z_1 и Z_2 можно эффективно воспользоваться методикой скоростных ортогональных преобразований, рассмотренной в предыдущем пункте. С этой целью следует в матрицах $\tilde{Z}_1 = Z_1 H$ и $\tilde{Z}_2 = Z_2 H$ найти столбцы с наибольшими нормами и тем самым задать соответственно шкалу конформизма и шкалу вариации для объектов.

Если же в матрицах $\tilde{Z}_1$ и $\tilde{Z}_2$ использовать по два столбца с наибольшими нормами, то получаются варианты двумерной визуализации системы объектов.

Предложенная методика изучения системы объектов A удобна в случае номинальных шкал и обладает хорошими скоростными качествами. Требуемое количество операций для получения картины взаимного расположения объектов системы A равно 2NM (преобразования $A \to Z_1$ или $A \to Z_2$) плюс $O(NM \lg_2 M)$ действий для быстрого преобразования Адамара или $O(NM)$ - для Хаара.

Итак, для номинальных шкал разработаны скоростные методы визуализации взаимного расположения объектов системы A. Как же быть с признаками? При решении вопроса быстрой визуализации взаимного расположения признаков возможны три подхода.

Первый подход заключается в использовании для матриц Z_1 и Z_2 скоростных ортогональных преобразований слева: $\tilde{Z}_1 = HZ_1$ и $\tilde{Z}_2 = HZ_2$.

Результатом преобразований (в случае необходимости уточнение производится методом частотной ортогонализации [3]) будет визуализация расположения признаков матрицы данных. Хотя этот метод и скоростной (порядка $NM \lg_2 N$), но для наших целей он слишком трудоемок, поскольку преобразованиям H подвергаются векторы-столбцы, содержащие большое число N элементов.

Второй подход, который дает одновременно двумерное расположение объектов и признаков, состоит в представлении матрицы данных в виде $\bar{A} = UT$. Матрица $U = N \times 2$ -матрица, а ее строки - пара координат; матрица $T - 2 \times M$ -матрица, ее столбцы - также пара координат. Разложение A вида UT требует всего 2NM операций.

При третьем подходе производится унификация матрицы A таким образом, чтобы гистограммы (на основании которых составляются частотная матрица и матрица вариации) можно было образовать симметрично как для объектов, так и для признаков.

По матрице номинальных данных A нельзя образовать гистограмму по объекту, так как признаки несоизмеримы, а градации, одинаково обозначенные у разных признаков несопоставимы. Заметим, что после проведения нормировки матрицы A при помощи преобразования $A \to Z_1$ или $A \to Z_2$, значения z_{ij} получают вероятностный смысл. Число z_{ij}/N - это оценка вероятности приобретения объектом значения a_{ij} в признаке j, иначе говоря, это вероятность данного состояния по данному признаку; l_j является нормирующим множителем для стандартизации вариации признака j.

Поэтому на основании преобразования $A \to Z_q = (1^q_j . z^q_{ij})$ можно предположить удобную стратегию унификации матрицы данных.

Стратегия состоит в следующем. Сначала все величины $1^q_j z^q_{ij}$, $i = 1,\ldots,N$; $j = 1,\ldots,M$, представляются на действительной оси в отрезке $(L = \min\limits_{i,j} 1^q_j z^q_{ij}$, $U = \max\limits_{i,j} 1^q_j . z^q_{ij})$. Затем этот отрезок разбивается на К однородных интервалов. Для такого разбиения можно воспользоваться какой-нибудь методикой кластерного анализа или же просто произвести равномерную разбивку отрезка, последовательно приписывая интервалам в направлении от L к U метки $I,2,\ldots,K$ (тем самым производим ординацию шкалы $[L, U]$). После ординации шкалы $[L, U]$ элементы матрицы A можно заменить метками $I,2,\ldots,K$.

Описанный здесь способ назовем унификацией матрицы A.

Элементы унифицированной матрицы обозначим через v_{ij}, а переход от A к матрице V через $A \to V$.

Таким образом, полученная матрица V унифицированных меток симметрична относительно объектов и признаков. По ней точно так, как и по схеме для признаков j, можно составить гистограммы объектов, вычислять конформности и вариации признаков и проверять их упорядочение.

На матрице унифицированных меток применимы все изложенные выше методы шкалирования и визуализации, а также все предложенные варианты быстрых ортогональных преобразований в форме левого и правого H-преобразований $\tilde{V}_q = V_q H$, $\bar{V}_q = HV_q$.

Наиболее важные ее применения можно найти в статье [4] о монотонных системах наблюдений.

Литература

I. ВЫХАНДУ Л.К. Экспресс-методы анализа данных. -Тр. Таллинского политехнического института. 1979, №464, с.21-37.

2. НИЛЬСОН А.А. Некоторые свойства сумм квадратов вероятностей и их математико-статистические приложения. -Изв. АН ЭССР, сер. техн. и физ.-мат.наук, 1965, т.14, с. 79-93.

3. ВЫХАНДУ Л.К. Некоторые проблемы теории анализа данных. - Тр. Таллинского политехнического института. 1974, № 366, с.3-14.

4. ВЫХАНДУ Л.К. О некоторых методах упорядочения объектов и признаков в системе данных.-Тр.Таллинского политехнического института. 1980, № 482, с.43-50.

5. BISHOP Y. e.a. Discrete multivariate analysis.-N.-Y.:1975.

Поступила в ред.-изд.отд.
28 февраля 1981 года