Энтропийный метод сжатия матриц со случайными значениями элементов[1]

Ю.С. Попков

Аннотация. Предлагается двухэтапный метод сжатия матрицы со случайными значениями элементов, имитирующей входные данные (прецеденты) в задачах рандомизированного машинного обучении (РМО). Ядро метода составляют операции <<прямого>> и <<обратного>> проектирования, оптимизация которых осуществляется посредством максимизации информационной относительной энтропии. Соответствующие этим операциям модули могут применяться в виде последовательной или параллельной процедуры. Разработаны алгоритмы максимизации информационной относительной энтропии, использующие покоординатную модификацию метода проекций градиента.

Ключевые слова: относительная информационная энтропия, операторы проектирования, матричные производные, градиентный метод, прямые и обратные проекции.

Введение

Процедуры рандомизированного машинного обучения (РМО) являются эффективным инструментальным средством решения задач классификации, регрессии и кластеризации в условиях неопределенности. Источниками ее являются модели и данные, входящие в состав обучающих коллекций. Обучающие коллекции включают две компоненты. Входные данные, характеризуемые матрицами определенных размеров $(m \times s)$, где m - количество объектов в обучающей коллекции, и s - размерность пространства признаков, в терминах которых характеризуются объекты. Выходные данные характеризуются m-мерным вектором, компонентами которого являются параметры принадлежности классу, либо параметры состояния объектов.

Основу концепции рандомизированного машинного обучения составляет гипотеза о случайной природе данных. Вообще говоря, обучающая коллекция (обе ее компоненты) содержит ошибки, но в процедурах обучения предполагается, что основные ошибки, порождающие неопределенность, сосредоточены в выходных данных.

По разным причинам в перечисленных выше задачах возникает необходимость <<сжать>> обучающую коллекцию, т.е. входную компоненту описывать $(n \times r)$-матрицей. Содержательно это сводится к редукции пространства признаков и уменьшению количества объектов, на массиве которых проводится обучение модели.

Данная проблема вложена в более общую: приближение заданного набора многомерных точек маломерным аффинным многообразием [4]. Наиболее распространенным для ее решения является метод главных компонент (МГК) [5, 6], его робастные версии [7] и многочисленные эмпирические модификации. Следует также отметить, что МГК представляет собой комбинацию математического метода получения упорядоченной последовательности собственных чисел матрицы данных и эвристических соображений по поводу <<обрезания>> этой последовательности.

[1]Работа поддержана РФФИ (гранты 17-07-00286, 17-29-03119).

Однако часто можно задать числа n, r. Здесь развиваются энтропийный метод сжатия матрицы данных и его реализации в виде последовательной и параллельной процедур. При последовательной процедуре сначала производится сжатие по размерности признакового пространства ($r < s$), а затем по количеству объектов во входной обучающей коллекции ($n < m$). При параллельной процедуре производится одновременное сжатие по обеим размерностям. Технологически в обеих методах реализуется процедура <<прямого>> и <<обратного>> проектирования с помощью матриц-<<проекторов>> соответствующих размеров. Поскольку метод сжатия ориентирован на приложения в РМО, то матрицы-<<проекторы>> определяются путем максимизации информационной относительной энтропии [10].

1. Энтропийное <<прямое>> и <<обратное>> проектирование

Рассмотрим пространство R^{NM} матриц $A_{(N\times M)}$ с вещественными элементами, которые принимают случайные значения [1]. Преобразуем матрицу $A_{N\times M}$ к матрице $B_{N\times m}$, где $m < M$. Это можно сделать с помощью двухшаговой процедуры. Сначала проектируем пространство R^{NM} на пространство R^{Nm} ($R^{NM} \Rightarrow R^{Nm}$) с помощью матрицы-<<проектора>> $S_{(N\times m)}$. Получаем матрицу $D_{(N\times m)} = A_{(N\times M)} S_{(M\times m)}$. Эту операцию будем назвать <<прямым>> проектированием.

Переход от пространства R^{NM} к пространству меньшей размерности R^{Nm} связан с потерей информации, заключенной в элементах исходного пространства. Для оценки этих потерь нужно вернуться в исходное пространство, т.е. осуществить операцию <<обратного>> проектирования: $R^{Nm} \Rightarrow R^{NM}$. Эта операция осуществляется с помощью матрицы-<<проектора>> $J_{(m\times M)}$. Получаем матрицу $X_{(N\times M)} = D_{(N\times m)} J_{(m\times M)} \in R^{NM}$. Матрицы $X_{(N\times M)}, A_{(N\times M)}$ являются элементами одного пространства, что дает возможность привлечь подходящий критерий оценки качества осуществляемого сжатия исходной матрицы.

Поскольку предполагается, что матрица $A_{(N\times M)}$ имеет элементы со случайными значениями, то целесообразно оценивать качество ее сжатия, которое зависит от матриц-<<проекторов>> $S_{(N\times m)}, J_{(m\times M)}$, в терминах энтропии, которая является характеристикой неопределенности, свойственной задачам РМО. Здесь будет использоваться информационная энтропия [10] матрицы $X_{(N\times M)}$ относительно реализации матрицы $A_{(N\times M)}$, которую далее будем называть информационной относительной энтропией:

$$H(X \mid A) = -\sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} ln \frac{x_{ij}}{a_{ij}} = tr(X^{\mathrm{T}} Y), \quad (1)$$

где tr - след матрицы, заключенной в скобках; Y - $(N \times M)$- матрица с элементами:

$$y_{ij} = \ln \frac{x_{ij}}{a_{ij}}. \quad (2)$$

Описанный метод <<прямого>> и <<обратного>> проектирования составляет ядро метода сжатия матрицы со случайными значениями элементов.

2. Последовательная процедура сжатия

Рассмотрим матрицу со случайным происхождением данных U размерности $(m \times s)$ и неотрицательными элементами. Ее нужно <<сжать>> до матрицы Z, размером $(n \times r)$, $r < s, n < m$. Будем решать эту задачу последовательно в два этапа. Сначала редуцируем матрицу $U_{(m\times s)}$ до матрицы $Y_{(m\times r)}$, а затем последнюю <<сожмем>> до матрицы $Z_{(n\times r)}$. Указанные операции будем осуществлять с помощью матриц-<<проекторов>>, упомянутых в предыдущем разделе.

1. *Первый этап*. Введем матрицу $Q_{(s\times r)}$ и осуществим операцию <<прямого>> проектирования для получения сжатой по количеству столбцов матрицы

$$Y_{(m\times r)} = U_{(m\times s)} Q_{(s\times r)}. \quad (3)$$

Введем матрицу $T_{(r\times s)}$ и осуществим операцию <<обратного>> проектирования. Получим матрицу

$$X_{(m\times s)} = U_{(m\times s)} Q_{(s\times r)} T_{(r\times s)}. \quad (4)$$

[1] При этом используется *векторизация* матрицы [12]

Матрица $X_{(m\times s)}$ имеет ту же размерность, что и исходная матрица $U_{(m\times s)}$, и является характеристикой последствий (потерь) операций <<прямого>> и <<обратного>> проектирования на первом этапе.

Редукция матрицы данных сопровождается потерей информации, которая заключена в исходной матрице, или увеличением относительной информационной энтропии [10]. Так, для первого этапа данного метода относительная информационная энтропия матрицы $X_{(m\times s)}$ при фиксированной исходной матрице $U_{(m\times s)}$ определяется следующим равенством:

$$H_1(X \mid U) = -\sum_{i=1}^{m} \sum_{j=1}^{s} x_{ij}(Q,T \mid U) \ln \frac{x_{ij}(Q,T \mid U)}{u_{ij}} = \\ = H_1(Q,T \mid U) \Rightarrow \max_{(Q,T)\geq 0}, \qquad (5)$$

где согласно (4)

$$x_{ij}(Q,T \mid U) = \sum_{\mu=1}^{r} \sum_{\nu=1}^{s} t_{\mu,j} q_{\nu,\mu} u_{i\nu}. \qquad (6)$$

В результате максимизации функции (5) получаем энтропийно-оптимальную матрицу-<<проектор>> $Q^*_{(s\times r)}$ и <<сжатую>> матрицу первого этапа

$$Y^*_{(m\times r)} = U_{(m\times s)} Q^*_{(s\times r)}. \qquad (7)$$

2. *Второй этап.* Матрица данных второго этапа $Y_{(m\times r)}$ имеет сокращенное количество столбцов $r < s$, и требуется сжать ее по количеству строк до заданной величины $n < m$. Введем матрицы $B_{(n\times m)}, A_{(m\times n)}$, с помощью которых реализуем снова операции <<прямого>> и <<обратного>> проектирования:

$$\begin{aligned} Z_{(n\times r)} &= B_{(n\times m)} Y_{(m\times r)}, \\ G_{(m\times r)} &= A_{(m\times n)} Z_{(n\times r)}. \end{aligned} \qquad (8)$$

Матрица $G_{(m\times r)}$ имеет ту же размерность, что и исходная матрица второго этапа $Y_{(m\times r)}$, и является характеристикой последствий (потерь) операций <<прямого>> и <<обратного>> проектирования на втором этапе. Используем информационную относительную энтропию для характеризации этих потерь:

$$H_2(G \mid Y^*) = -\sum_{i=1}^{m} \sum_{j=1}^{r} g_{ij}(B,A \mid Y) \ln \frac{g_{ij}(B,A \mid Y^*)}{y_{ij}} = \\ = H_2(B,A \mid Y^*) \Rightarrow \max_{(B,A)\geq 0}, \qquad (9)$$

где

$$g_{ij}(B,A \mid Y^*) = \sum_{\beta=1}^{n} \sum_{\alpha=1}^{m} a_{i,\beta} b_{\beta,\alpha} y^*_{\alpha,j}. \qquad (10)$$

$$y^*_{\alpha,j} = \sum_{\eta=1}^{s} u_{\alpha,\eta} q^*_{\eta,j}. \qquad (11)$$

3. Энтропийная оптимизация последовательной процедуры

Итак, первый этап последовательной процедуры состоит в определении оптимальных элементов неотрицательных матриц Q, T:

$$(Q^*, T^*) = \arg \max_{(Q,T)\geq 0} H_1(Q,T \mid U). \qquad (12)$$

Численное решение этой задачи будем осуществляеть с помощью итерационного алгоритма, основанного на методе проекций градиента (комбинация градиентного спуска и отсечения отрицательных элементов).

Компоненты градиента функции $H_1(Q,T \mid U)$ по элементам матриц Q, T имеют следующий вид:

$$\frac{\partial H(Q,T \mid U)}{\partial q_{\nu,\mu}} = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\partial h_{ij}(Q,T \mid U)}{\partial x_{ij}} \frac{\partial x_{ij}(Q,T \mid U)}{\partial q_{\nu,\mu}}, \qquad (13)$$
$$\nu = \overline{1,s}, \mu = \overline{1,r},$$

$$\frac{\partial H(Q,T \mid U)}{\partial t_{\mu,j}} = \sum_{i=1}^{s} \sum_{j=1}^{m} \frac{\partial h_{ij}(Q,T \mid U)}{\partial x_{ij}} \frac{\partial x_{ij}(Q,T \mid U)}{\partial t_{\mu,j}}, \qquad (14)$$
$$\mu = \overline{1,r}, j = \overline{1,s},$$

где

$$h_{ij}(T,Q \mid X) = x_{ij} \ln \frac{x_{ij}}{u_{ij}},$$
$$\frac{\partial h_{ij}(T,Q \mid X)}{\partial x_{ij}} = \ln \frac{x_{ij}}{u_{ij}} + 1, \quad i = \overline{1,m}, j = \overline{1,s}; \qquad (15)$$

$$\frac{\partial x_{ij}((Q,T \mid U))}{\partial q_{\nu,\mu}} = t_{\mu,j}\, u_{i,\nu}, \quad \nu = \overline{1,s}, \mu = \overline{1,r}; \qquad (16)$$
$$\frac{\partial x_{ij}((Q,T \mid U))}{\partial t_{\mu,j}} = \sum_{\nu=1}^{s} q_{\nu,\mu} u_{i,\nu}, \quad \mu = \overline{1,r}; \qquad (17)$$
$$i = \overline{1,m}, \quad j = \overline{1,s}.$$

Произведем векторизацию [12] матриц Q, T и введем соответствующие векторы q, t. Обозначим $\nabla_Q(q,t)$ - вектор градиента функционала относительной энтропии $H(Q,T \mid U)$ с компонентами (15),(16), $\nabla_T(q,t)$ - вектор градиента функционала относительной энтропии $H(Q,T \mid U)$ с компонентами (15),(17).

Алгоритм $\mathbf{max}\ \boldsymbol{H_1}$ имеет вид:

а). начальный шаг

$$q^0 > 0, \quad t^0 > 0;$$

б).n-ый итерационный шаг

$$X^n = U\, Q^n\, T^n;$$

$$H(Q^n, T^n \mid U) = \sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}^n \ln\frac{x_{ij}^n}{u_{ij}} = H^n.$$

$$q^{(n+1)} = \begin{cases} (q^n + \gamma_q \nabla_Q(q^n, t^n), \text{если } q^{(n+1)} \geq 0, \\ q^n, \text{если } q^{(n+1)} > 0 \end{cases}$$

$$q^{(n+1)} \Rightarrow Q^{(n+1)};$$

$$t^{(n+1)} = \begin{cases} t^n + \gamma_t \nabla_T(q^n, t^n), \text{если } t^{(n+1)} \geq 0, \\ t^n, \text{если } t^{(n+1)} < 0. \end{cases}$$

$$t^{(n+1)} \Rightarrow T^{(n+1)};$$

$$X^{(n+1)} = U\, Q^{(n+1)}\, T^{(n+1)};$$

$$H(Q^{(n+1)}, T^{(n+1)} \mid U) = \sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}^{(n+1)} \ln\frac{x_{ij}^{(n+1)}}{u_{ij}} = \\ = H^{(n+1)}.$$

в). Условие остановки.

$$\text{если } H^{(n+1)} - H^{(n)} \leq \delta, \qquad \text{то } STOP.$$

Аппроксимация функционала относительной энтропии. Иногда может оказаться достаточным для определения элементов редуцированной матрицы Q приближенное выражение - функционала, построенное на аппроксимации логарифмической функции в окрестности точки $x_0 = a$, где $a \geq a_{min}$:

$$\ln x < \ln a + \frac{1}{a_{min}}(x - a).$$

Применяя эту аппроксимацию для $a = x_{ij}$, получим следующее приближенное выражение:

$$H_1(Q, T \mid U) \approx \\ \approx \sum_{i=1}^{s}\sum_{j=1}^{m}\left[x_{ij}^2(Q, T \mid U) - x_{ij}(Q, T \mid U)u_{ij}\right] = \\ = \widetilde{H}(Q, T \mid U). \tag{18}$$

Представим это равенство в матричной форме:

$$\widetilde{H}_1(Q, T \mid U) = tr(XX^T) - tr(XU^T) = \\ = \lceil X(Q, T), X(Q, T) \rfloor - \lceil X(Q, T), U, \rfloor \tag{19}$$

где $\lceil A, B \rfloor$ - фробениусово скалярное произведение [12],

$$tr(AB^T) = tr(BA^T) = \lceil A, B \rfloor = \lceil B, A \rfloor. \tag{20}$$

Элементы матриц Q, T будем определять минимизируя функционал $\widetilde{H}_1(Q, T \mid U)$ на множестве неотрицательных матриц Q, T при фиксированной матрице данных U, т.е.

$$(\tilde{Q}^*, \tilde{T}^*) = \arg\min_{(Q,T)\geq 0} \widetilde{H}_1(Q, T \mid U). \tag{21}$$

Алгоритм решения этой задачи использует компоненты градиента функционала $\widetilde{H}_1(Q, T \mid U)$ по матрицам Q и T.

Напомним правила дифференцирования некоторых скалярных функций, зависящих от матриц [12]. Рассмотрим линейное пространство $(m \times n)$-матриц A. Скалярное произведение в этом пространстве:

$$\lceil A, B \rfloor = tr(AB^T) = tr(BA^T).$$

Норма (фробениусова):

$$\parallel A \parallel^2 = \lceil A, A \rfloor.$$

Из этих определений видно, что матрица рассматривается как вектор, составленный из примыкающих строк (*векторизация* матрицы). Для скалярной функции от матрицы градиент определяется следующим образом. Рассмотрим скалярную функцию $f(A)$. Ее приращение представимо в виде:

$$f(A + \Delta) = f(A) + \lceil H, \Delta \rfloor + \Delta$$

Здесь Δ - остаток ряда Тейлора. Градиент функции $f(A)$ по матрице A

$$\Delta_A f(A) = H.$$

Приведем простые, но полезные примеры скалярных функций от матрицы:

$$f(A) = \lceil C, A \rfloor, \qquad \Delta_A f(A) = C.$$

$$f(A) = \parallel A \parallel^2, \qquad \Delta_A f(A) = 2A.$$

Используя правила матричного дифференцирования, изложенные выше, получим следующие выражения для градиентов функционала $\widetilde{H}_1(Q, T \mid U)$ (21):

$$\Delta_Q(Q, T) = \frac{\partial \widetilde{H}_1(Q, T \mid U)}{\partial X}\frac{\partial X}{\partial Q} = 2TPQT - TP. \tag{22}$$

$$\Delta_T(Q, T) = \frac{\partial \widetilde{H}_1(Q, T \mid U)}{\partial X}\frac{\partial X}{\partial T} = 2Q^T PQT - Q^T P, \tag{23}$$

Здесь $P = XX^T$. В этих равенствах приняты следующие обозначения: $\Delta_Q(Q, T), \Delta_T(Q, T)$ - градиенты по матрицам Q и T, соответственно.

Согласно (22, 23) градиенты функции $\widetilde{H}(Q,T \mid U)$ по матрицам Q и T представляет собой матрицы размером $(s \times r)$.

Для максимизации функционала $\widetilde{H}_1(Q,T \mid U)$ воспользуемся градиентным спуском с контролем условий неотрицательности матриц U, T:

Алгоритм $\mathbf{max}\ \widetilde{\boldsymbol{H}}_{\mathbf{1}}$ имеет вид:

а). начальный шаг

$$T^0 > 0, \quad Q^0 > 0;$$

б).n-ый итерационный шаг

$$X^n = U\, Q^n\, T^n;$$

$$H(Q^n, T^n \mid U) = \sum_{i=1}^{m} \sum_{j=1}^{r} x_{ij}^n \ln \frac{x_{ij}^n}{u_{ij}} = H^n.$$

$$Q^{(n+1)} = \begin{cases} Q^n + \gamma_Q \Delta_Q\left(\widetilde{H}(Q^n, T^n \mid U\right), \text{если } Q^{(n+1)} \geq 0, \\ Q^n, \text{если } Q^{(n+1)} < 0. \end{cases}$$

$$T^{(n+1)} = \begin{cases} T^n + \gamma_T \Delta_T(\widetilde{H}(Q^n, T^n \mid U), \text{если } T^{(n+1)} \geq 0, \\ T^n, \text{если } T^{(n+1)} < 0. \end{cases}$$

$$X^{(n+1)} = U\, Q^{(n+1)}\, T^{(n+1)};$$

$$H(Q^{(n+1)}, T^{(n+1)} \mid U) = \\ = \sum_{i=1}^{m} \sum_{j=1}^{r} x_{ij}^{(n+1)} \ln \frac{x_{ij}^{(n+1)}}{u_{ij}} = H^{(n+1)}.$$

Условие остановки.

Если $H^{(n+1)} - H^n \leq \delta$, то $STOP$.

В результате применения алгоритмов $\mathbf{max}\ \boldsymbol{H}_{\mathbf{1}}$ или $\mathbf{max}\ \widetilde{\boldsymbol{H}}_{\mathbf{1}}$ получаем энтропийно-оптимальную матрицу-<<проектор>> $Q^*_{(s\times r)}$. Следовательно, можем определить матрицу $Y^*_{(m\times r)}$ (7) для перехода ко второму этапу сжатия матрицы U, который характеризуется матрицами-<<проекторами>> $B_{(n\times m)}$ и $A_{(m\times n)}$.

Оптимальные элементы неотрицательных матриц B, A определяются следующим выражением:

$$(B^*, A^*) = \arg \max_{(B,A)\geq 0} H_2(B, A \mid Y^*). \tag{24}$$

Численное решение этой задачи будем осуществлять также, как и на первом этапе, с помощью итерационного алгоритма, основанного на методе проекций градиента (комбинация градиентного спуска и отсечения отрицательных элементов).

Компоненты градиента функции $H_2(B, A \mid Y^*)$ по элементам матриц B, A имеют следующий вид:

$$\frac{\partial H_2(B, A \mid Y^*)}{\partial a_{i,\beta}} = \\ = \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{\partial s_{ij}(B, A \mid Y^*)}{\partial g_{ij}} \frac{\partial g_{ij}(B, A \mid Y^*)}{\partial a_{i,\beta}}, \\ i = \overline{1,n};\ j = \overline{1,r};\ \beta = \overline{1,n}, \tag{25}$$

$$\frac{\partial H_2(B, A \mid Y^*)}{\partial b_{\beta,\alpha}} = \\ = \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{\partial s_{ij}(B, A \mid Y^*)}{\partial g_{ij}} \frac{\partial g_{ij}(B, A \mid Y^*)}{\partial b_{\beta,\alpha}}, \\ i = \overline{1,n};\ j = \overline{1,r};\ \beta = \overline{1,n};\ \alpha = \overline{1,m}. \tag{26}$$

где

$$\frac{\partial s_{ij}(B,A \mid Y^*)}{\partial g_{ij}} = \ln \frac{g_{ij}}{y_{ij}^*} + 1, \quad i = \overline{1,n};\ j = \overline{1,r}; \tag{27}$$

$$\frac{\partial g_{ij}((B,A \mid Y^*))}{\partial a_{i,\beta}} = \sum_{\alpha=1}^{m} b_{\beta,\alpha} y_{\alpha,j}^*, \qquad \beta = \overline{1,n}; \tag{28}$$

$$\frac{\partial g_{ij}((B,A \mid Y^*))}{\partial b_{\beta,\alpha}} = a_{i,\beta} y_{\alpha,j}^*, \quad \beta = \overline{1,n}; \alpha = \overline{1,m}; \tag{29}$$

$$i = \overline{1,m}; \quad j = \overline{1,r}.$$

Произведем векторизацию матриц B, A и введем соответствующие векторы b, a. Обозначим $\nabla_B(b, a)$ - вектор градиента функционала относительной энтропии $H_2(B, A \mid Y^*)$ с компонентами (27),(28), $\nabla_A(q, t)$ - вектор градиента функционала относительной энтропии $H_2(B, A \mid Y^*)$ с компонентами (27),(29).

Алгоритм $\mathbf{max}\ \boldsymbol{H}_{\mathbf{2}}$ имеет вид:

а) начальный шаг

$$b^0 > 0, \quad a^0 > 0;$$

б) n-ый итерационный шаг

$$G^n = A^n B^n Y^*;$$

$$H(B^n, A^n \mid Y^*) = \sum_{i=1}^{n} \sum_{j=1}^{r} g_{ij}^n \ln \frac{g_{ij}^n}{y_{ij}^*} = H^n.$$

$$b^{(n+1)} = \begin{cases} b^n + \gamma_b \nabla_B(b^n, a^n), \text{если } b^{(n+1)} \geq 0 \\ b^n, \text{ если } b^{(n+1)} < 0. \end{cases}$$

$$b^{(n+1)} \Rightarrow Q^{(n+1)};$$

$$a^{(n+1)} = \begin{cases} a^n + \gamma_a \nabla_A\left(b^n, a^n\right), \text{если } a^{(n+1)} \geq 0, \\ a^n, \text{ если } a^{(n+1)} < 0. \end{cases}$$

$$a^{(n+1)} \Rightarrow A^{(n+1)};$$

$$G^{(n+1)} = A^{(n+1)} B^{(n+1)} Y^*;$$

$$H(B^{(n+1)}, A^{(n+1)} \mid Y^*) = \\ = \sum_{i=1}^{n} \sum_{j=1}^{r} g_{ij}^{(n+1)} \ln \frac{g_{ij}^{(n+1)}}{y_{ij}^*} = H^{(n+1)}.$$

в). Условие остановки.

Если $H^{(n+1)} - H^{(n)} \leq \delta$, то $STOP$.

5. Энтропийная оптимизация параллельной процедуры

Параллельная реализация <<прямого>> и <<обратного>> проектирования приводит к следующим матричным равенствам:

$$U_{(m\times s)} Q_{s\times r} = Y_{(m\times r)}, \quad B_{(n\times m)} Y_{(m\times r)} = Z_{(n\times r)}, \\ Z_{(n\times r)} W_{(r\times s)} = D_{(n\times s)}, \quad E_{(m\times n)} D_{(n\times s)} = X_{(m\times s)}. \quad (29)$$

Первая цепочка <<прямой>> проекции с матрицей $Q_{(s\times r)}$ и <<обратной>> проекции с матрицей $B_{(n\times m)}$ дает редуцированную матрицу $Z_{(n\times r)}$, а вторая - с матрицей $W_{(r\times s)}$ <<прямой>> проекции и с матрицей $E_{(m\times n)}$ <<обратной>> проекции дает матрицу $X_{(m\times s)}$.

Из равенств (29) имеем:

$$X_{(m\times s)} = E_{(m\times n)}\left\{\left[B_{(n\times m)}\left(U_{(m\times s)} Q_{(s\times r)}\right)\right] W_{(r\times s)}\right\}. \quad (30)$$

Скобки в этом равенстве указывают на последовательность операций проектирования. Элементы матрицы $X_{(m\times s)}$ имеют вид:

$$x_{ij} = \sum_{\mu=1}^{n} e_{i,\mu} \sum_{\nu=1}^{r} w_{\nu,j} \sum_{\beta=1}^{m} b_{\mu,\beta} \sum_{\alpha=1}^{s} u_{\beta,\alpha} q_{\alpha,\nu}, \\ i = \overline{1,m}, j = \overline{1,s}. \quad (31)$$

Для измерения отклонения преобразованной матрицы $X_{(m\times s)}$ от исходной $U_{(m\times s)}$ воспользуемся относительной энтропией

$$\mathcal{H}(X \mid U) = -\sum_{i=1}^{m} \sum_{j=1}^{s} s_{ij}(X \mid U), \quad (32)$$

где

$$s_{ij} = x_{ij} \ln \frac{x_{ij}}{u_{ij}}. \quad (33)$$

С учетом равенства (31) не трудно видеть, что относительная энтропия (32) есть скалярная функция от матриц $(Q, B, W, E) \geq 0$, т.е.

$$\mathcal{H} = \mathcal{H}(Q, B, W, E). \quad (34)$$

Оптимальные значения элементов этих матриц определяются максимумом функции $\mathcal{H}$:

$$(Q^*, B^*, W^*, E^*) = \arg \max_{(Q,B,W,E)\geq 0} \mathcal{H}(Q, B, W, E). \quad (35)$$

Для численного решения этой задачи воспользуемся покоординатной схемой метода проекций градиента.

В параллельной процедуре выделяется пара матриц Q, B, участвующие в редукции исходной матрицы U по одной переменной ($r < s$), и пара матриц W, E, с помощью которых осуществляется <<сжатие>> матрицы U по второй переменной ($n < m$). Итерационный шаг покоординатной схемы метода проекций градиента состоит из двух подшагов: на одном осуществляется итерация по (Q, B)-градиентам, а на другом - по (W, E)-градиентам.

Согласно (31 - 33) компоненты градиента функции $\mathcal{H}(Q, B, W, E)$ имеют вид:

$$\frac{\partial \mathcal{H}}{\partial q_{\alpha,\nu}} = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\partial s_{ij}}{\partial x_{ij}} \frac{\partial x_{ij}}{\partial q_{\alpha,\nu}},$$

$$\frac{\partial \mathcal{H}}{\partial b_{\mu,\beta}} = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\partial s_{ij}}{\partial x_{ij}} \frac{\partial x_{ij}}{\partial b_{\mu,\beta}},$$

$$\frac{\partial \mathcal{H}}{\partial w_{\nu,j}} = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\partial s_{ij}}{\partial x_{ij}} \frac{\partial x_{ij}}{\partial w_{\nu,j}},$$

$$\frac{\partial \mathcal{H}}{\partial e_{i,\mu}} = \sum_{i=1}^{m} \sum_{j=1}^{s} \frac{\partial s_{ij}}{\partial x_{ij}} \frac{\partial x_{ij}}{\partial e_{i,\mu}}, \quad (36)$$

где

$$\frac{\partial s_{ij}}{\partial x_{ij}} = \ln \frac{x_{ij}}{u_{ij}} + 1,$$

$$\frac{\partial x_{ij}}{\partial q_{\alpha,\nu}} = w_{\nu,j} \sum_{\mu=1}^{n} e_{i,\mu} \sum_{\beta=1}^{m} b_{\mu,\beta} u_{\beta,\alpha}, \\ \alpha = \overline{1,s};\ \nu = \overline{1,r};$$

$$\frac{\partial x_{ij}}{\partial b_{\mu,\beta}} = e_{i,\mu} \sum_{\nu=1}^{r} w_{\nu,j} \sum_{\alpha=1}^{s} u_{\beta,\alpha} q_{\alpha,\nu}, \\ \mu = \overline{1,n};\ \beta = \overline{1,s};$$

$$\frac{\partial x_{ij}}{\partial w_{\nu,j}} = \sum_{\mu=1}^{n} e_{i,\mu} \sum_{\beta=1}^{m} b_{\mu,\beta} \sum_{\alpha=1}^{s} u_{\beta,\alpha} q_{\alpha,\nu}, \\ \nu = \overline{1,r};\ j = \overline{1,s};$$

$$\frac{\partial x_{ij}}{\partial e_{i,\mu}} = \sum_{\nu=1}^{r} w_{\nu,j} \sum_{\beta=1}^{m} b_{\mu,\beta} \sum_{\alpha=1}^{s} u_{\beta,\alpha} q_{\alpha,\nu}, \\ i = \overline{1,m};\ \mu = \overline{1,n}; \quad (37)$$

Применяя векторизацию матриц, перейдем к соответствующим векторам q, b, w, e. Обозначим компоненты градиента по этим векторам $\nabla_q, \nabla_b, \nabla_w, \nabla_e$.

Алгоритм максимизации max $\boldsymbol{\mathcal{H}}$ имеет следующий вид:

а) начальный шаг

$$q^0 > 0, \quad b^0 > 0, \quad w^0 > 0, \quad e^0 > 0;$$

б) n-ый итерационный шаг

$$X^n = E^n\{[B^n(UQ^n)]W^n\}$$

$$\mathcal{H}^n = \sum_{i=1}^{m}\sum_{j=1}^{s} x_{ij}^n \ln\frac{x_{ij}^n}{u_{ij}}.$$

$$q^{(n+1)} = \begin{cases} q^n + \gamma_q \nabla_q(q^n, b^n, w^n, e^n), \text{если } q^{(n+1)} \geq 0, \\ q^n, \text{ если } q^{(n+1)} < 0. \end{cases}$$

$$q^{(n+1)} \Rightarrow Q^{(n+1)};$$

$$b^{(n+1)} = \begin{cases} b^n + \gamma_b \nabla_b(q^n, b^n, w^n, e^n), \text{если } b^{(n+1)} \geq 0, \\ b^n, \text{ если } b^{(n+1)} < 0. \end{cases}$$

$$b^{(n+1)} \Rightarrow B^{(n+1)};$$

$$w^{(n+1)} = \\ = \begin{cases} w^n + \gamma_w \nabla_w(q^{(n+1)}, b^{(n+1)}, w^n, e^n), \text{если } w^{(n+1)} \geq 0, \\ w^n, \text{ если } w^{(n+1)} < 0. \end{cases}$$

$$w^{(n+1)} \Rightarrow W^{(n+1)};$$

$$e^{(n+1)} = \\ = \begin{cases} e^n + \gamma_e \nabla_e(q^{(n+1)}, b^{(n+1)}, w^n, e^n), \text{если } e^{(n+1)} \geq 0, \\ e^n, \text{ если } e^{(n+1)} < 0. \end{cases}$$

$$e^{(n+1)} \Rightarrow E^{(n+1)};$$

$$X^{(n+1)} = E^{(n+1)}\{\left[B^{(n+1)}\left(UQ^{(n+1)}\right)\right]W^{(n+1)}\};$$

$$\mathcal{H}^{(n+1)} = \sum_{i=1}^{m}\sum_{j=1}^{s} x_{ij}^{(n+1)} \ln\frac{x_{ij}^{(n+1)}}{u_{ij}};$$

в) условие остановки.

Если $\mathcal{H}^{(n+1)} - \mathcal{H}^{(n)} \leq \delta, \qquad$ то $STOP.$

Заключение

Предлагается двухэтапный метод сжатия матрицы со случайными значениями элементов, имитирующей входные данные (прецеденты) в задачах рандомизированного машинного обучении (РМО). Ядро метода составляют операции <<прямого>> и <<обратного>> проектирования, оптимизация которых осуществляется посредством максимизации информационной относительной энтропии. Соответствующие этим операциям модули могут применяться в виде последовательной или параллельной процедуры. Разработаны алгоритмы максимизации информационной относительной энтропии, использующие покоординатную модификацию метода проекций градиента.

Литература

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М.. Фининсы и статистика, 1989.
2. Мерков А.Б. Распознавание образов. Введение в методы статистического обучения. М., Едиториал УРСС, 2010.
3. Имельбаев Ш.С., Шмульян Б.Л. Моделирование стохастических коммуникационных систем.// Вильсон А.Дж. Энтропийные методы моделирования сложных систем. М., Наука, 1978, с.170-234.
4. Bruckstein A.M., Donoho D.L., Elad M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Rev. 2009, v.51, No.1, p.34-81.
5. Кендал М.Дж., Стюарт А. Статистические выводы и связи./ Пер. с англ. М.. Наука, 1973.
6. Jollife I.T. Principal Component Analysis. N.Y. Springer-Verlag, 2002.
7. Поляк Б.Т., Хлебников М.В. Метод главных компонент: робастные версии. Автоматика и Телемеханика, 2017, №3, с.130-148.
8. Воронцов К.В. Математические методы обучения по прецендентам. Курс лекций МФТИ, 2006.
9. Jaynes E.T. Information Theory and Statistical Mechanics. Physics Review, 1957, v.106, p. 620-639.
10. Popkov Yu.S. Macrosystem Theory and Applications. Springer, Lecture Notes in Informatics and Control, 1995.
11. Kullback S., Leibler R.A. On information and Sufficiency. Ann. of Math. Statistics, 1951, v. 22(1), p. 79-86.
12. Magnus J., Neudecker H. Matrix Differential Calculus with Applications in Statistics and Econometrics. John Willey and Sons, 2007, thied Edt.

Попков Юрий Соломонович. Директор ИСА ФИЦ ИУ РАН. Окончил МЭИ в 1960 году. Академик, доктор технических наук, профессор. Количество печатных работ: 200. Область научных интересов: математическое программирование, динамические макросистемы. E-mail: popkov.yuri@gmail.com

Entropy methad of cjvpresion matrices with random elements

Popkov Yu.S.

Two-stages method of compresion matrices with random elements is proposed. Operation of direct and invers projections are the kernal of method. The information cross-entropy is used for optimization these operations. Serial and paralel procedures are proposed.

Keywords: cross-entropy, projection operator, matrix derivatives, gradien method.

References

1. Aivazian S.A, Buchshtaber V.M., Eniukov I.S., Meshalkin L.D. Prikladnaia statistika: klassifikacia i snizhenie razmernosti. M.: Finansi I statistika, 1989.
2. Merkov A.B. Raspoznavanie obrazov. Vvedenie v metodi statisticheskogo obuchenia. M. Editorial URSS, 2010.
3. Imelbaev S.S., Shmulian B.L. Modelirovie stochasticheskih kommunikacionnih sistem. M. Nauka, 1978, c.170-234.
4. Bruckstein A.M., Donoho D.L., Elad M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Rev. 2009, v.51, No.1, p.34-81.
5. Kendal M.J., Stewart A. Statisticheskie vivody I sviazi. M. Nauka, 1973.
6. Jollife I.T. Principal Component Analysis. N.Y. Springer-Verlag, 2002.
7. Polyak B.T., Hlebnikov M.V. Metod glavnih komponent: robastnie versii // Avtomatika I telemehanika, 2017, №3, c.130-148.
8. Vorontsov K.V. Matematicheskie metodi obuchenia po precedentam. Kurs lekcii MFTI, 2006.
9. Jaynes E.T. Information Theory and Statistical Mechanics. Physics Review, 1957, v.106, p. 620-639.
10. Popkov Yu.S. Macrosystem Theory and Applications. Springer, Lecture Notes in Informatics and Control, 1995.
11. Kullback S., Leibler R.A. On information and Sufficiency. Ann. of Math. Statistics, 1951, v.22(1), p. 79-86.
12. Magnus J., Neudecker H. Matrix Differential Calculus with Applications in Statistics and Econometrics. John Willey and Sons, 2007, thied Edt.

Popkov Yuri Solomonovich. Director of ISA FRC CSC RAS. D.Sc, Professor of Braude College, Karmiel, Israel. The number of publications: 200 publications. Research interests: mathematial programming, dynamic macrosystems.
E-mail: popkov.yuri@gmail.com