Нечеткая линейная регрессия при горизонтально распределенных данных

Е. С. Волкова[1], В. Б. Гисин[2]
Финансовый университет при Правительстве Российской Федерации
[1]evolkova@fa.ru, [2]vgisin@fa.ru

Аннотация. **Основная идея коллаборативного построения общей модели состоит в том, что несколько клиентов сотрудничают для построения общей модели. Модель строится относительно совокупного набора данных. При этом данные каждого клиента должны оставаться конфиденциальными. В работе рассматривается задача коллаборативного построения нечеткой линейной регрессии при горизонтальном разделении данных. В этом случае регрессоры и объясняемая переменная являются общими для всех клиентов. В то же время каждый клиент обладает своим набором наблюдений, содержащих значения регрессоров и объясняемой переменной. Предложено решение, основанное на трансформационном подходе.**

Ключевые слова: нечеткая линейная регрессия; конфиденциальность; коллаборативные вычисления; горизонтально распределеннные данные; трансформационный подход

I. Введение

Современное развитие информационных технологий опирается в значительной степени на машинное обучение и обработку больших данных. В ряде случаев задача обучения осложняется тем, что необходимые для этого данные хранятся у нескольких клиентов. Если должна быть сохранена конфиденциальность данных каждого клиента, традиционные модели машинного обучения приходится заменять специальными протоколами.

Например, при построении линейной (или логистической) регрессии данные наблюдений и значения зависимой величины могут быть распределены по нескольким клиентам. В случае, когда данные распределены горизонтально, т. е. все клиенты имеют один и тот же набор регрессоров и общую зависимую величину, для решения задачи построения общей регрессионной модели разработан ряд протоколов федеративного машинного обучения [1].

Как известно [2, 3], построение нечеткой линейной регрессии сводится к задаче линейного программирования. Для решения задач линейного программирования с горизонтальным распределением данных применяется трансформационный подход, при котором клиенты кодируют свои данные и решают задачу, используя интерактивный протокол. Этот подход был описан в работах [5, 6] и затем усовершенствован в работах [7–9].

В случае, когда среди ограничений в задаче линейного программирования имеются неравенства, введение балансовых переменных приводит к проблемам с обеспечением конфиденциальности данных [10, 11]. В задачах линейного программирования, возникающих при построении нечеткой линейной регрессии, используется большое число балансовых переменных. Это требует дополнительных мер маскировки исходной информации клиентов.

В [12] описан протокол конфиденциального двустороннего вычисления параметров нечеткой линейной регрессионной модели. В настоящей работе описывается протокол построения нечеткой линейной регрессии с числом участников, большим, чем два.

II. Нечеткая линейная регрессия

Задача построения нечеткой линейной регрессии ставится следующим образом. Имеется объясняемая величина y и набор регрессоров x_j, $j = 0,\ldots, n$. Предполагается, что величина y представляет собой линейную комбинацию регрессоров с нечеткими коэффициентами. Требуется определить вектор коэффициентов $A = (A_0,\ldots,A_n)^t$, имея m наблюдений $((x_{ij}), y_i)$ $i = 1,\ldots, m$ (считается, что $x_{i0} = 1$ для всех i). Коэффициенты нужно определить так, чтобы для каждой нечеткой величины

$$Y_i = A_0 + A_1x_{i1} +\ldots+ A_nx_{in}$$

выполнялось соотношение $\mu_i(y_i) \ge h$, $i = 1,\ldots, m$, где μ_i — функция принадлежности величины Y_i, а h — некоторое заданное заранее пороговое значение.

Иными словами, если $X = (x_{ij})$ и $Y = XA$, то $\mu_Y(y) \ge h$, где $y = (y_1,\ldots,y_m)^t$, а принадлежность к векторной нечеткой величине определяется покоординатно с помощью оператора min. При этом суммарная нечеткость коэффициентов должна быть минимальной.

Следуя [3], мы предполагаем, что все коэффициенты A_j — симметричные треугольные числа, которые заданы своим центральным значением a_j и спрэдом $\alpha_j \ge 0$. В этом случае Y_i — также треугольное число с центром

$$a_0 + a_1x_{i1} +\ldots+ a_nx_{in}$$

и спрэдом

$$(1 - h)(\alpha_0 + \alpha_1|x_{i1}| +\ldots+ \alpha_n|x_{in}|).$$

Обозначим через $a = (a_j)^t$ вектор центральных значений, а через $\alpha = (\alpha_j)^t$ вектор спрэдов. Тогда условие $\mu_Y(y) \ge h$ запишется в виде

$$Xa - (1 - h)|X|\alpha \le y \le Xa + (1 - h)|X|\alpha, \qquad (1)$$

где $|X| = (|x_{ij}|)$.

Минимальная суммарная нечеткость коэффициентов определяется как

$$\min (|X|^t I_m)^t\alpha, \qquad (2)$$

где I_m — вектор размерности $m > 0$, все координаты которого равны единице.

Тем самым задача построения нечеткой линейной регрессии сводится к задаче линейного программирования, в которой требуется найти минимальное значение (2) при выполнении ограничений (1) и условии $\alpha \geq 0$.

В приложениях, основываясь на содержательных соображениях, можно указать такой вектор d, размерности $n + 1$, что $a + d \geq 0$. С учетом этого можно условие $\alpha \geq 0$ заменить условием неотрицательности всех переменных: $a \geq 0$, $\alpha \geq 0$.

Положим

$$J(X) = \begin{pmatrix} X & -(1-h)\,|\,X\,| \\ -X & -(1-h)\,|\,X\,| \end{pmatrix}, \; j(y) = (y, -y)^t.$$

Тогда с учетом сделанного замечания задача определения коэффициентов нечеткой линейной регрессии приобретает следующий вид:

$$(|X|^t\, I_m)^t \alpha \to \min;\; J(X)(a, \alpha)^t \leq j(y);\; a, \alpha \geq 0.$$

Вводя вектор балансовых переменных размерности $2m$, приходим к следующей задаче:

$$(|X|^t\, I_m)^t \alpha \to \min;\; (J(X)\;\; E_{2m})z = j(y);\; z \geq 0, \qquad (3)$$

где z — вектор, составленный из вектора $(a, \alpha)^t$, дополненного вектором балансовых переменных размерности m, а E_{2m} — единичная матрица размерности $2m$.

Задача (3), как нетрудно заметить, всегда имеет решения. При горизонтальном распределении данных строки матрицы коэффициентов $J(X)$ распределены между клиентами. Каждый клиент k располагает своей матрицей $J^{(k)}(X)$. Протокол коллаборативного вычисления позволяет собрать данные клиентов и получить задачу (3) в трансформированном виде так, чтобы сохранить конфиденциальность данных каждого клиента.

Считается, что размерностные параметры доступны всем клиентам. Клиенты являются полу-честными (честными, но любопытными), т. е., честно исполняют протокол, но стремятся извлечь из получаемых по протоколу данных дополнительную информацию об исходных данных других участников.

III. Трансформация

Пусть p — число участников (клиентов), занумерованных числами от 1 до p. Обозначим через m_k — число наблюдений, принадлежащих клиенту k. Тогда $m = m_1 + \ldots + m_p$. Будем считать, что наблюдения клиента k представлены матрицей $X^{(k)}$ и вектором $y^{(k)}$ так, как это описано в предыдущем разделе.

На подготовительном этапе каждый из участников k генерирует случайный вектор $\theta^{(k)}$ размерности $2m_k$ и случайные обратимые матрицы $P^{(k)}$ и $Q^{(k)}$ размерности $2m_k$.

Ограничения вида (3) в форме уравнений преобразуются следующим образом:

$$J^{(k)} s^{(k)} = r^{(k)}, \qquad (4)$$

где

$$J^{(k)} = \begin{pmatrix} J(X^{(k)}) & E_{2m_k} & P^{(k)} \\ 0 & 0 & Q^{(k)} \end{pmatrix}; \; r^{(k)} = \begin{pmatrix} j(y^{(k)}) + P^{(k)}\theta^{(k)} \\ Q^{(k)}\theta^{(k)} \end{pmatrix},$$

а $s^{(k)}$ — вектор переменных, в котором первые $2n + 2$ координаты отведены под переменные $(a, \alpha)^t$, следующие $2m_k$ позиции — под балансовые переменные и последние $2m_k$ позиций — под фиктивные переменные, значения которых составляют вектор $Q^{(k)}\theta^{(k)}$.

В целом задача (3) может быть сформулирована теперь следующим образом:

$$c^t s \to \min,\; T(X)s = r,\; s \geq 0. \qquad (5)$$

Здесь:

$r = (r^{(1)}, \ldots, r^{(p)})^t$;

координаты вектора c — это координаты вектора $(|X|^t\, I_m)^t$, дополненные соответствующим числом нулей;

матрица $T(X)$ имеют следующую блочную структуру:

$$T(X) = (J\;\; D),$$

где блочный столбец J сформирован из вертикально соединенных клеток вида

$$\begin{pmatrix} J(X^{(k)}) \\ 0 \end{pmatrix}, \; k = 1, \ldots, p,$$

а D — блочно-диагональная матрица, диагональные клетки которой имеют вид

$$\begin{pmatrix} E_{2m_k} & P^{(k)} \\ 0 & Q^{(k)} \end{pmatrix}.$$

Высота матрицы $T(X)$ составляет $4m$, ширина — $2(n + 1) + 4m$.

В ходе исполнения первой фазы протокола обмена информацией между клиентами должно быть сформировано уравнение вида

$$KT(X)Mw = Kr, \qquad (6)$$

где K, M — обратимые матрицы, известные лишь клиенту p, а само уравнение (6) формируется у клиента 1.

Очевидным образом устанавливается соответствие между решениями уравнения (6) относительно w и решениями уравнения из (5). Достаточно заметить, что вектор Mw является решением уравнения $T(X)s = r$ тогда и только тогда, когда вектор w является решением уравнения (6).

Формирование уравнения (6) происходит по следующей схеме. Информация об уравнении вида (4) последовательно передается в зашифрованном виде от первого клиента второму, от второго — третьему и т. д.

Шифр представляет собой массив матриц. Клиент p генерирует обратимые матрицы K и M и выполняет преобразование KHM для всех матриц H, полученных им от клиента $p - 1$, направляет результат клиенту $p - 1$. Клиент $p - 1$, в свою очередь, проводит частичное декодирование и пересылает результаты клиенту $p - 2$.

Этот процесс продолжается, пока информация не достигнет клиента 1. Проведя частичное декодирование, клиент 1 получает матрицу коэффициентов уравнения (6).

Формирование правой части уравнения (6) происходит достаточно просто. Каждый из клиентов $k = 1, \ldots, p-1$ пересылают столбец $r^{(k)}$ клиенту p. Клиент p формирует столбец r, умножает его слева на известную ему матрицу K и направляет результат клиенту 1. Маскировка исходных векторов $j(y^{(k)})$ обеспечивается их суммированием со случайным вектором $P^{(k)}\theta^{(k)}$.

Для выполнения шифрования при формировании матрицы $KT(X)M$ клиенты генерируют свои секретные ключи. Ключ представляет собой случайное число $\omega < 2^q$, где q — параметр безопасности. Пусть $\omega(t)$ — двоичная цифра на месте t в двоичном разложении числа ω, так что

$$\omega = \omega(1)\, 2^{q-1} + \ldots + \omega(q).$$

Далее, обозначим через $F^{(k)}$ матрицу того же размера, что и матрица $T(X)$, такую, что соответствующие $4m_k$ строк этой матрицы совпадают со строками матрицы $T(X)$, а остальные элементы равны нулю.

Клиент 1 генерирует случайные матрицы $H(1), \ldots, H(q-1)$ и матрицу $H(q)$ так, что

$$H(1) + \ldots + H(q-1) + H(q) = F^{(1)}.$$

Далее, используя свой ключ $\omega^{(1)}$, клиент 1 генерирует случайные матрицы $G^{(1)}(t, 1 - \omega^{(1)}(t))$, $t = 1, \ldots, q$, и пересылает клиенту 2 массив матриц

$$\mathbf{G^{(1)}}[1{:}q;0{:}1],$$

в котором

$$\mathbf{G^{(1)}}(t, \omega^{(1)}(t)) = H(t),$$

$$\mathbf{G^{(1)}}(t, 1 - \omega^{(1)}(t)) = G^{(1)}(t, 1 - \omega^{(1)}(t))$$

при всех $t = 1, \ldots, q$.

Клиент 2 аналогичным образом формирует матрицы

$$G^{(2)}(t,\tau),\ t = 1, \ldots, q,\ \tau = 0, 1.$$

Затем он пересылает клиенту 3 массив матриц

$$\mathbf{G^{(2)}}[1{:}q;\ 0{:}1;\ 1{:}q;\ 0{:}1]$$

такой, что

$$\mathbf{G^{(2)}}(t_1,\tau_1;\ t_2,\tau_2) = G^{(1)}(t_1,\tau_1) + G^{(2)}(t_2,\tau_2).$$

Аналогичные действия выполняют клиенты до $p-1$-го.

Клиент p начинает с того, что формирует матрицу $F^{(p)}$. Положим $G^{(p)} = F^{(p)}$. Далее, используя сгенерированные им обратимые матрицы K и M, клиент p формирует массив матриц

$$\mathbf{T}^{(p)}[1{:}q;\ 0{:}1;\ldots;1{:}q;\ 0{:}1]$$

такой, что

$$\mathbf{T}^{(p)}(t_1,\tau_1;\ldots;\ t_{p-1},\tau_{p-1}) =$$

$$= K(G^{(1)}(t_1,\tau_1) + \ldots + G^{(p-1)}(t_{p-1},\tau_{p-1}) + G^{(p)})M, \qquad (7)$$

$$t_{1,2,\ldots,p-1} = 1, \ldots, q,\ \tau_{1,2,\ldots,p-1} = 0, 1.$$

Этот массив клиент p пересылает клиенту $p-1$.

Клиент $p-1$ проводит частичное декодирование, суммируя матрицы массива $\mathbf{T}^{(p)}$ индексами последнего уровня t_{p-1}, $\omega^{(p-1)}(t_{p-1})$ при t_{p-1}, пробегающим значения от 1 до q и фиксированных остальных индексах. Таким образом, клиент $p-1$ вычисляет

$$\sum_{t_{p-1}=1}^{q} \mathbf{T}^{(p)}(t_1, \tau_1;\ldots;t_{p-1}, \omega^{(p-1)}(t_{p-1}))$$

при всех

$$t_1, t_2,\ldots, t_{p-2} = 1, \ldots, q, \tau_1, \tau_1, \ldots, \tau_{p-2}, = 0, 1.$$

При этом он получает массив матриц $\mathbf{T}^{(p-1)}$, который отправляет клиенту $p-2$, и т.д.

Как несложно убедиться, массив $\mathbf{T}^{(1)}$ состоит из одной матрицы $q^{p-2}KT(X)M$:

$$\mathbf{T}^{(1)} = \{q^{p-2}KT(X)M\}.$$

Таким образом, у клиента 1 оказывается сформированной матрица $KT(X)M$.

Аналогичная процедура используется для формирования векторов целевой функции

$$c^tMw \to \min$$

(ключи $\omega^{(k)}$, естественно, формируются заново).

Линейной трансформации подвергаются также условия неотрицательности. Если неотрицательность старых переменных непосредственно свести к неотрицательности новых переменных, матрица M должна быть мономиальной, т. е. иметь в каждой строке и в каждом столбце ровно один ненулевой элемент, и все ее элементы должны быть при этом неотрицательны [6].

Как показано в [10], такой подход не гарантирует приемлемого уровня конфиденциальности. Заметим также, что использование вместо линейного преобразования $s = Mw$ аффинного преобразования также не дает дополнительных гарантий защищенности информации [11]. Для формирования условий неотрицательности мы воспользуемся идеей из [7], заменив аффинное преобразование линейным и введя некоторые дополнительные меры защиты.

Для трансформации условий неотрицательности из (5) клиент p генерирует случайную квадратную матрицу L порядка $4m$ такую, что

$$LKr = 0.$$

Для дополнительной маскировки используется положительная мономиальная матрица B порядка $2(n+1) + 4m$, которую формирует клиент p.

Трансформированные условия неотрицательности могут быть записаны при этом следующим образом:

$$(BM - LKT(X)M)w \geq 0.$$

Процедура формирования матрицы $BM - LKT(X)M$ у клиента 1 происходит по той же схеме, что и формирование матрицы $KT(X)M$.

В итоге клиент 1 получает следующую задачу линейного программирования:

$$c^t Mw \to \min;\ KT(X)Mw = Kr,\ (B - LKT(X))Mw \geq 0. \quad (8)$$

IV. Корректность и конфиденциальность

Во-первых, покажем, что задачи (5) и (8) эквивалентны.

В самом деле, пусть s — решение задачи (5). Положим $w = M^{-1}s$. Тогда

$$KT(X)Mw = KT(X)s = Kr.$$

Далее,

$$(B - LKT(X))Mw = Bs - LKT(X)s =$$

$$= Bs - LKr = Bs \geq 0,$$

поскольку $s \geq 0$, а матрица B мономиальная.

Наконец, целевая функция $c^t Mw = c^t s$ принимает минимальное значение, поскольку множества допустимых решений в задачах (5) и (8) находятся во взаимно однозначном соответствии.

Обратно, пусть w — решение задачи (8). Для $s = Mw$ имеем $T(X)s = r$, поскольку $KT(X)Mw = Kr$, а матрица K обратима. Далее, неравенство из (8) равносильно тому, что $Bs \geq 0$. Но это последнее условие означает, что $s \geq 0$ с учетом того, что B — мономиальная матрица.

Решив задачу (8), клиент 1 высылает решение w клиенту p. Клиент p вычисляет вектор $s = Mw$. Первые $2(n + 1)$ координат этого вектора дают центры и спрэды искомых нечетких коэффициентов.

Задачи линейного программирования, возникающие при построении нечеткой линейной регрессии, имеют специфическую структуру (дублирование матрицы X и матрицы ее абсолютных значений, большое число балансовых переменных, вид коэффициентов целевой функции). С учетом этого в протоколе предусмотрены дополнительные меры защиты информации по сравнению, с другими протоколами решения задач линейного программирования: введены дополнительные ключевые объекты (матрицы $P^{(k)}$, $Q^{(k)}$ и вектор $\theta^{(k)}$, ключи $\omega^{(k)}$), использованы специфические приемы разделения информации.

В [12] приведено обоснование защищенности протокола двустороннего конфиденциального вычисления параметров нечеткой линейной регрессии. В его основе лежит тот факт, что матрицы и векторы в трансформированной задаче могут быть произвольными. Общими для них с векторами и матрицами исходной задачи служат лишь размерностные характеристики (включая ранг). Это справедливо и для протокола, описанного в настоящей работе. Сохранение конфиденциальности в ходе интерактивного взаимодействия обеспечивается использованием ключей $\omega^{(k)}$. Заметим, впрочем, что длина ключа (параметр безопасности q) является для описанного протокола достаточно критичной, поскольку объем передаваемой информации в процессе взаимодействия клиентов экспоненциально зависит от длины ключа.

V. Заключение

В работе приводится описание протокола коллаборативного вычисления параметров нечеткой линейной регрессии при горизонтальном разделении данных. Решение опирается на трансформационный подход, модифицированный с учетом специфической структуры задач линейного программирования, возникающих при построении нечеткой линейной регрессии. Дальнейшие исследования могут быть связаны с поиском принципиально иных решений, основанных на идеях федеративного машинного обучения.

Список литературы

[1] Yang Q., Liu Y., Chen T., Tong Y. Federated machine learning: Concept and applications // ACM Transactions on Intelligent Systems and Technology (TIST). 2019. Т. 10, №. 2. С. 1-19.

[2] Tanaka H., Uejima S., Asai K. Linear regression analysis with fuzzy model // IEEE Transactions on Systems, Man, and Cybernetics. 1982. Т. 12, №. 6. С. 903-907.

[3] Tanaka H. Fuzzy data analysis by possibilistic linear models // Fuzzy sets and systems. 1987. Т. 24, № 3. С. 363-375.

[4] Tanaka H., Watada J. Possibilistic linear systems and their application to the linear regression model // Fuzzy sets and systems. 1988. Т. 27, №. 3. С. 275-289.

[5] Mangasarian O. L. Privacy-preserving linear programming // Optimization Letters. 2011. Т. 5, №. 1. С. 165-172.

[6] Mangasarian O. L. Privacy-preserving horizontally partitioned linear programs // Optimization Letters. 2012. Т. 6, №. 3. С. 431-436.

[7] Wang C., Ren K., Wang J. Secure optimization computation outsourcing in cloud computing: A case study of linear programming //IEEE transactions on computers. 2016. v. 65. №. 1. pp. 216-229.

[8] Secure Outsourcing of Large-scale Linear Programming / Z. Wang Z., L. I. U. Yang // In: 2017 2nd International Conference on Wireless Communication and Network Engineering (WCNE 2017) / DEStech Transactions on Computer Science and Engineering. 2017. №. wcne. pp. 185-190 DOI:10.12783/dtcse/wcne2017/19821

[9] Hong Y. Vaidya, J., Rizzo, N., & Liu, Q. Privacy-preserving linear programming // World scientific reference on innovation: Volume 4: Innovation in Information Security. 2018. pp. 71-93. https://doi.org/10.1142/9789813149106_0004

[10] On the (Im)possibility of privately outsourcing linear programming / P. Laud, A. Pankova // Proceedings of the 2013 ACM workshop on Cloud computing security workshop. 2013. С. 55-64.

[11] New Attacks against Transformation-Based Privacy-Preserving Linear Programming/ P.Laud, A. Pankova //International Workshop on Security and Trust Management. Springer, Berlin, Heidelberg, 2013. С. 17-32.

[12] Волкова Е.С., Гисин В.Б. Конфиденциальное двустороннее вычисление параметров нечеткой линейной регрессионной модели //Вопросы кибербезопасности. 2021. №. 3. С. 11-19.