Вестник Сыктывкарского университета.
Серия 1: Математика. Механика. Информатика. 2022.
Выпуск 1 (42)
Bulletin of Syktyvkar University.
Series 1: Mathematics. Mechanics. Informatics. 2022; 1 (42)

ПРИКЛАДНАЯ МАТЕМАТИКА И МЕХАНИКА

Научная статья

УДК 519.8
https://doi.org/10.34130/1992-2752_2022_1_15

КВАДРАТИЧНАЯ ЗАДАЧА МАТЕМАТИЧЕСКОЙ ДИАГНОСТИКИ

Надежда Олеговна Котелина[1], Александр Борисович Певный[2]
[1,2]Сыктывкарский государственный университет им. Питирима Сорокина, [1]nkotelina@gmail.com, [2]pevnyi@syktsu.ru

Аннотация. Пусть в n-мерном пространстве даны m точек, а G – выпуклая оболочка этих точек. В простейшей задаче математической диагностики проверяется, принадлежит ли точка p множеству G. Иначе говоря, если координаты точек – признаки некоторой болезни, необходимо определить, есть ли у нового пациента болезнь по степени схожести ее признаков у него и у пациентов с подтвержденным диагнозом. В настоящей статье мы присоединяем к G его эпсилон-окрестность и проверяем, принадлежит ли p расширенному множеству. Для этого решаем задачу квадратичного программирования, в которой требуется найти точку множества G, ближайшую к точке p в евклидовой норме. Мы выписываем необходимые условия минимума, получая задачу, которую можно решать при помощи модифицированного симплекс-метода с дополнительным условием для базисов.

Ключевые слова: математическая диагностика, машинное обучение, модифицированный симплекс-метод, квадратичное программирование

Для цитирования: Котелина Н. О., Певный А. Б. Квадратичная задача математической диагностики // *Вестник Сыктывкарского университета. Сер. 1: Математика. Механика. Информатика.* 2022. Вып. 1 (42). С. 15–22. https://doi.org/10.34130/1992-2752_2022_1_15

Applied mathematics and mechanics

Original article

Quadratic problem of mathematical diagnostics

Nadezhda O. Kotelina, Aleksandr B. Pevnyi
[1,2]Pitirim Sorokin Syktyvkar State University, [1]nkotelina@gmail.com, [2]pevnyi@syktsu.ru

Abstract. Let m points be given in n-dimensional space, and G is the convex hull of these points. In the simplest problem of mathematical diagnostics, it is checked whether a point p belongs to the set G. In other words, if the coordinates of the points are signs of some disease, it is necessary to determine whether a new patient has a disease by the similarity of its signs in him and in patients with a confirmed diagnosis. In this paper, we attach its epsilon neighborhood to G and check whether p belongs to an extended set. To do this, we solve a quadratic programming problem in which we need to find the point of the set G closest to the point p in the Euclidean norm. In the article, we write out the necessary minimum conditions, obtaining a problem that can be solved using a modified simplex method with an additional condition for the bases.

Keywords: mathematical diagnostics, machine learning, modified simplex-method, quadratic programming

For citation: Kotelina N. O., Pevnyi A. B. Quadratic problem of mathematical diagnostics. *Bulletin of Syktyvkar University, Series 1: Mathematics. Mechanics. Informatics*, 2022, No. 1 (42), pp. 15−22. https://doi.org/10.34130/1992-2752_2022_1_15

Введение

Пусть в пространстве $\mathbb{R}^n$ задано конечное множество точек $\{p_j\}_{j=1}^m$. Каждый вектор p_j состоит из n компонент – это признаки (факторы) некоторой болезни у больных с подтвержденным диагнозом. Есть ещё точка p. Есть ли у этого нового пациента болезнь, определяется по степени похожести признаков точки p и признаков точек $\{p_j\}$.

В простейшей задаче математической диагностики [1] вводится выпуклая оболочка $G = \mathrm{co}\,\{p_1, \ldots, p_m\}$ точек $\{p_j\}$ и проверяется включение $p \in G$.

Было бы интересно присоединить к G точки, у которых расстояние до G мало. Зададим $\varepsilon > 0$. Найдём расстояние от p до G:

$$\rho = \min_{z \in G} \|p - z\|, \tag{1}$$

где $\|p-z\|$ – евклидова норма $p-z$. Если $\rho \le \varepsilon$, то считаем, что диагноз (тест) положительный.

Данная статья посвящена численному решению задачи (1).

Материалы и методы

Введём векторы $a_j = p_j - p$ и выпуклую оболочку этих векторов:

$$M = \mathrm{co}\,\{a_1, \ldots, a_m\}.$$

Требуется найти расстояние от многогранника M до начала координат:

$$\rho = \min_{z \in M} \|z\|. \tag{2}$$

Здесь и далее используется евклидова норма $\|z\| = \sqrt{(z, z)}$ и обычное скалярное произведение $(z,\, y)$ векторов $z,\, y \in \mathbb{R}^n$.

Множество M состоит из векторов $z = \sum_{j=1}^{m} x_j a_j$, где $x_j \ge 0$, $x_1 + \cdots + x_m = 1$. Поэтому задача (2) эквивалентна задаче квадратичного программирования

$$\begin{aligned} f(x) = \tfrac{1}{2}(Dx,\, x) &\to \min, \\ x_1 + \cdots + x_m &= 1, \\ x_j \ge 0, &\quad j \in 1:m, \end{aligned} \tag{3}$$

где D – матрица с элементами $d_{ij} = (a_i,\, a_j)$, $i,\, j \in 1:m$.

Запишем необходимые условия минимума в задаче (3) – условия Куна-Таккера. Градиент $f'(x)$ равен Dx, а градиент $g(x) = x_1 + \cdots + x_m$ есть $g'(x) = \mathbb{I}$ – вектор из m единиц. Пусть $(Dx)_i$ – i-я компонента вектора Dx. В точке минимума x^* существует λ, такое, что

$$(Dx^*)_i + \lambda \ge 0, \quad i \in 1:m, \tag{4}$$

$$((Dx^*)_i + \lambda) \cdot x_i^* = 0, \quad i \in 1:m. \tag{5}$$

Смысл условия дополняющей нежёсткости (5) состоит в том, что если $x_i^* > 0$, то $(Dx^*)_i + \lambda = 0$. Условия (4) – (5) являются и достаточными условиями минимума, так как матрица D неотрицательно определена.

Чтобы найти x^*, введём дополнительные переменные $v_i = (Dx)_i + \lambda$. Тогда получим систему

$$\begin{aligned} &(Dx)_i + \lambda - v_i = 0, \quad i \in 1:m, \\ &x_1 + \dots + x_m = 1, \\ &v_i x_i = 0, \quad v_i \geq 0,\ x_i \geq 0,\ i \in 1:m. \end{aligned} \tag{6}$$

Несмотря на нелинейное ограничение $v_i x_i = 0$, здесь можно применить идеи линейного программирования. В первое уравнение добавим искусственную переменную u со знаком минус. Придём к задаче:

$$\begin{gathered} u \to \min, \\ Dx + \lambda \cdot \mathbb{I} - v - e_1 u = \mathbb{O}, \\ x_1 + \dots + x_m = 1, \\ v_i x_i = 0, \quad v_i \geq 0,\ x_i \geq 0,\ i \in 1:m;\ u \geq 0. \end{gathered} \tag{7}$$

Здесь $e_1 = (1, 0, ..., 0)^T$, $\mathbb{I} = (1, 1, ..., 1)^T$.

Задача (7) решается модифицированным симплекс-методом с дополнительным условием: если x_i входит в базис, то v_i нельзя вводить в базис, и наоборот. Использовать только базисы с таким свойством дополнительности предложил П. Вулф [3]. Алгоритм с этой старой идеей описан в статье [2]. Некоторая сложность возникает при определении начального базиса.

В задаче (6) переменная λ принимает значения любого знака. Поэтому представим λ в виде $\lambda = \lambda_1 - \lambda_2$, где $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. Для наглядности выпишем систему уравнений при $m = 4$:

$$\begin{aligned} &d_{11}x_1 + d_{12}x_2 + d_{13}x_3 + d_{14}x_4 + \lambda_1 - \lambda_2 - v_1 - u &&= 0, \\ &d_{21}x_1 + d_{22}x_2 + d_{23}x_3 + d_{24}x_4 + \lambda_1 - \lambda_2 - v_2 &&= 0, \\ &d_{31}x_1 + d_{32}x_2 + d_{33}x_3 + d_{34}x_4 + \lambda_1 - \lambda_2 - v_3 &&= 0, \\ &d_{41}x_1 + d_{42}x_2 + d_{43}x_3 + d_{44}x_4 + \lambda_1 - \lambda_2 - v_4 &&= 0, \\ &x_1 + x_2 + x_3 + x_4 &&= 1. \end{aligned}$$

В первом столбце матрицы D найдём наименьший элемент

$$d_{i_0 1} = \min_{i \in 1:m} d_{i1}.$$

Если $i_0 = 1$, то $(a_i, a_1) \geq (a_1, a_1)$, а тогда $(z, a_1) \geq (a_1, a_1)$ и $(z - a_1, a_1) \geq 0$ для всех $z \in M$. Отсюда

$$\|z\|^2 = \|a_1 + (z - a_1)\|^2 = \|a_1\|^2 + 2(z - a_1, a_1) + \|z - a_1\|^2 \geq \|a_1\|^2.$$

Значит, a_1 – ближайшая точка, $z^* = a_1$, $\rho = \|a_1\|$ и задача (2) решена.

В дальнейшем считаем $i_0 > 1$. Укажем начальный базисный план. Идею построения покажем для частного случая $i_0 = 3$. Базисными переменными будут

$$u = d_{11} - d_{i_0 1},\ v_2 = d_{21} - d_{i_0 1},\ \lambda = -d_{i_0 1},\ v_4 = d_{41} - d_{i_0 1},\ x_1 = 1.$$

Если $d_{i_0 1} \geq 0$, то берём $\lambda_2 = d_{i_0 1}$, иначе $\lambda_1 = -d_{i_0 1} = |d_{i_0 1}|$. При $d_{i_0 1} \geq 0$ базисная матрица A и обратная базисная матрица $B = A^{-1}$ будут иметь вид

$$A = \begin{bmatrix} -1 & 0 & -1 & 0 & d_{11} \\ 0 & -1 & -1 & 0 & d_{21} \\ 0 & 0 & -1 & 0 & d_{i_0 1} \\ 0 & 0 & -1 & -1 & d_{41} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} -1 & 0 & 1 & 0 & d_{11} - d_{i_0 1} \\ 0 & -1 & 1 & 0 & d_{21} - d_{i_0 1} \\ 0 & 0 & -1 & 0 & d_{i_0 1} \\ 0 & 0 & 1 & -1 & d_{41} - d_{i_0 1} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (8)$$

Легко проверить, что $AB = E$.

Если же $d_{i_0 1} < 0$, то в матрице A вместо столбца $-\mathbb{I}$ будет столбец $+\mathbb{I}$, а в матрице B у элементов $b_{i_0 i_0}$ и $b_{i_0, m+1}$ меняются знаки.

В компьютерной программе матрица A не нужна, а будет использоваться только матрица B. Отметим также, что элементы $(m+1)$-го столбца B – это значения базисных переменных. Столбец X базисных значений получается из уравнения $AX = e_{m+1} = (0, 0, \ldots, 0, 1)^T$. Отсюда $X = Be_{m+1}$ – последний столбец матрицы B.

Чтобы обеспечить выполнение условия $x_i v_i = 0$ введём булевский массив $xv[1..m]$, где $xv[i] = true$, если x_i или v_i входят в базис.

Если $xv[i] = false$, то x_i и v_i не входят в базис и тогда вычисляем оценки

$$\Delta(x_i) = \sum_{k=1}^{m} y[k] \cdot d_{ki} + y[m+1], \quad (9)$$

$$\Delta(v_i) = -y[i], \quad (10)$$

где y – двойственный вектор. На всех итерациях строка c из коэффициентов целевой функции при базисных переменных имеет вид $c = (1, 0, \ldots, 0)$. Поэтому $y = cB$ – это первая строка матрицы B. Массив y можно не вводить.

Если λ_1 и λ_2 не входят в базис, то вычисляем оценку

$$\Delta(\lambda_{1,2}) = \pm \sum_{k=1}^{m} B[1, k].$$

На нулевой итерации в базис можно вводить только x_{i_0} или v_{i_0}. Первая строка матрицы B указана в формуле (8). По формуле (10) получаем $\Delta(v_{i_0}) = -B[1,\, i_0] = -1$. А по формуле (9)

$$\begin{aligned}\Delta(x_{i_0}) = -d_{1i_0} + d_{i_0 i_0} + (d_{11} - d_{i_0 1}) = \\ = \|a_{i_0}\|^2 - 2(a_{i_0},\, a_1) + \|a_1\|^2 = \|a_{i_0} - a_1\|^2.\end{aligned}$$

Если $a_{i_0} \neq a_1$, то x_{i_0} вводится в базис. Если же $a_{i_0} = a_1$, то и процесс заканчивается (нет положительных оценок). При этом

$$u = d_{11} - d_{i_0 1} = (a_1,\, a_1) - (a_1, a_{i_0}) = 0.$$

Условия (6) выполняются.

Остальные шаги модифицированного симплекс-метода подробно описаны в докладе [1].

Результаты

В 2017 году в Сыктывкарском государственном университете имени Питирима Сорокина была выполнена дипломная работа, где описанный метод реализован на C++. В работе был, в частности, такой пример: $m = n = 100$, $a_j = e_j$, $j \in 1 : n$. Ближайшей к началу координат точкой многогранника M будет $z^* = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. Программа выдала $z^* = (0.01, 0.01, \ldots, 0.01)$ на 99-й итерации.

Обсуждение

Решение простейшей задачи математической диагностики при помощи модифицированного симплекс-метода обсуждалось в докладе В. Н. Малозёмова и Е. К. Чернэуцану на семинаре по оптимизации, машинному обучению и искусственному интеллекту [1]. Данная статья развивает идею этого доклада, при этом для решения полученной задачи квадратичного программирования используется модифицированный симплекс-метод, идея которого восходит к статье П. Вулфа [3].

Список источников

1. **Малозёмов В. Н., Чернэуцану Е. К.** Простейшая задача математической диагностики // *Семинар «O & ML». Избранные доклады. 9 февраля 2022 г.* URL: http://www.apmath.spbu/oml/reps22.shtml#0209 (дата обращения: 04.04.2022).

2. **Певный А. Б.** Нахождение точки многогранника, ближайшей к началу координат // *Оптимизация.* Новосибирск, 1972. Вып. 10 (4).

3. **Wolfe P.** The simplex method for quadratic programming // *Econometrics.* 1959. Vol. 27. Pp. 382–398.

References

1. **Malozemov V. N., Cherneutsanu E. K.** The simplest problem of mathematical diagnostics. *Seminar «O & ML». Izbrannye doklady* [Seminar «O & ML». Selected papers]. 2022-02-09. Available: http://www.apmath.spbu/oml/reps22.shtml#0209 (accessed: 04.04.2022).

2. **Pevnyi A. B.** Finding the point of polyhedron closest to the origin (in Russian). *Optimizaciya* [Optimization]. Issue 10 (4). Novosibirsk, 1972.

3. **Wolfe P.** The simplex method for quadratic programming. *Econometrics.* 1959. Vol. 27. Pp. 382–398.

Сведения об авторах / Information about authors
Надежда Олеговна Котелина / Nadezhda O. Kotelina
к.ф.-м.н., доцент кафедры прикладной математики и информационных технологий в образовании / Ph.D. in Physics and Mathematics, Associate Professor of Department of Applied Mathematics and Computer Science
Сыктывкарский государственный университет им. Питирима Сорокина / Pitirim Sorokin Syktyvkar State University
167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 167001, Russia, Syktyvkar, Oktyabrsky Ave., 55

Александр Борисович Певный / Aleksandr B. Pevnyi
д.ф.-м.н., профессор, кафедра прикладной математики и информационных технологий в образовании / Doctor of physical and mathematical sciences, professor, Professor of Department of Applied Mathematics and Computer Science
Сыктывкарский государственный университет им. Питирима Сорокина / Pitirim Sorokin Syktyvkar State University
167001, Россия, г. Сыктывкар, Октябрьский пр., 55 / 167001, Russia, Syktyvkar, Oktyabrsky Ave., 55

Статья поступила в редакцию / The article was submitted 01.03.2022
Одобрено после рецензирования / Approved after reviewing 11.03.2022
Принято к публикации / Accepted for publication 15.03.2022