ОБ ОДНОЙ ЗАДАЧЕ РАСПОЗНАВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ, ВКЛЮЧАЮЩЕЙ ПОВТОРЯЮЩИЙСЯ ВЕКТОР[1]

Алексей Долгушев, Александр Кельманов

***Аннотация**: Рассматривается дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторных последовательностей, включающих в качестве элемента квазипериодически повторяющийся вектор евклидова пространства. Обоснован эффективный алгоритм решения задачи, гарантирующий оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределённых случайных величин.*

***Ключевые слова**: помехоустойчивое распознавание, векторная последовательность, повторяющийся вектор, максимум правдоподобия, дискретная экстремальная задача, off-line алгоритм.*

***ACM Classification Keywords**: F.2. Analysis of Algorithms and Problem Complexity, G.1.6. Optimization, G2. Discrete Mathematics, I.5. Pattern Recognition.*

***Conference**: The paper is selected from International Conference "Classification, Forecasting, Data Mining" CFDM 2009, Varna, Bulgaria, June-July 2009*

Введение

Объект исследования работы – проблемы оптимизации в задачах анализа данных и распознавания образов. Предмет исследования – дискретная экстремальная задача, к которой сводится один из вариантов проблемы помехоустойчивого off-line распознавания векторной последовательности, как последовательности, включающей квазипериодически повторяющийся вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Цель работы – обоснование алгоритма решения этой задачи. Рассматриваемая задача является обобщением задачи, изученной в [1].

Одна из возможных содержательных трактовок задачи состоит в следующем. Источник сообщений через канал связи с помехой передает информацию об активном и пассивном состояниях некоторого физического объекта в виде упорядоченного набора – вектора – измеряемых характеристик. В пассивном состоянии значения каждой компоненты этого вектора равны нулю. Имеется конечная совокупность физических объектов. Каждому объекту соответствует уникальный набор измеряемых информационно важных характеристик. На приёмную сторону поступает зашумлённая последовательность квазипериодически перемежающихся векторов, в которой кроме информационно значимого вектора, соответствующего активному состоянию объекта, имеются посторонние неизвестные ненулевые векторы-вставки. Термин «квазипериодически» означает, что интервал между двумя последовательными ненулевыми векторами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Число повторов информационно значимого вектора, а также число векторов-вставок известны. Требуется определить (распознать), от какого из объектов была принята последовательность. Ситуации, в которых требуется решение подобной задачи, характерны, в частности, для геофизики, технической диагностики, электронной разведки и других приложений (см., например, [2] и цитированные там работы).

[1] Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

Формальная постановка задачи

Пусть векторная последовательность $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1,2,\ldots,N\}$, обладает свойством

$$\mathbf{x}_n = \begin{cases} \mathbf{u}, & n \in \mathcal{M}_1, \\ \mathbf{w}_n, & n \in \mathcal{M}_2, \\ \mathbf{0}, & n \in \mathcal{N} \setminus (\mathcal{M}_1 \cup \mathcal{M}_2), \end{cases} \quad (1)$$

где $\mathcal{M}_1 \cup \mathcal{M}_2 \subseteq \mathcal{N}$, $\mathcal{M}_1 \cap \mathcal{M}_2 = \varnothing$.

Допустим, что $\mathbf{u} \in \mathcal{A}$, $\mathcal{A} \subset \{\mathbf{u} \,|\, \mathbf{u} \in \mathcal{R}^q, 0 < \|\mathbf{u}\|^2 < \infty\}$, где $\|\cdot\|$ – норма вектора, и $|\mathcal{A}| = K$. Пусть $\mathbf{w}_n \in \{\mathbf{w} \,|\, \mathbf{w} \in \mathcal{R}^q, 0 < \|\mathbf{w}\|^2 < \infty\}$, $n \in \mathcal{M}_2$.

Положим $|\mathcal{M}_j| = M_j$, $j = 1, 2$, и $M = M_1 + M_2$. Вектор $\mathbf{u}$ будем интерпретировать как информационно значимый вектор, множество $\mathcal{A}$ – как алфавит информационно значимых векторов, вектор $\mathbf{w}_n$, $n \in \mathcal{M}_2$, – как вектор-вставку, а M_1 и M_2 – соответственно как число повторов информационно значимого вектора и число векторов-вставок в последовательности (1). Допустим, кроме того, что элементы набора $(n_1,\ldots,n_M)$, образующего совокупность $\{n_1,\ldots,n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$, удовлетворяют ограничениям

$$1 \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N-1, \quad m = 2,\ldots,N. \quad (2)$$

Ограничения (2), в которых $T_{\min}$ и $T_{\max}$ – константы, задают допустимый интервал между ближайшими номерами двух ненулевых векторов последовательности (1).

Доступной для анализа будем считать последовательность

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathcal{N}, \quad (3)$$

где $\mathbf{e}_n$ – вектор помехи (ошибки измерения), независимый от вектора $\mathbf{x}_n$. Заметим, что $\mathbf{x}_n = \mathbf{x}_n(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}, \{\mathbf{w}_n, n \in \mathcal{M}_2\})$, $n \in \mathcal{N}$. Положим

$$S(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}, \{\mathbf{w}_n, n \in \mathcal{M}_2\}) = \sum_{n \in \mathcal{N}} \|\mathbf{y}_n - \mathbf{x}_n\|^2 \quad (4)$$

и рассмотрим следующую задачу.

Задача 1. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, и множество (алфавит) $\mathcal{A}$, $|\mathcal{A}| = K$. *Найти*: вектор $\mathbf{u} \in \mathcal{A}$, непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N}$, а также множество $\{\mathbf{w}_n \,|\, n \in \mathcal{M}_2, \mathbf{w}_n \in \mathcal{R}^q, 0 < \|\mathbf{w}_n\|^2 < \infty\}$ такие, что целевая функция (4) минимальна при ограничениях (2) на элементы набора $(n_1,\ldots,n_M)$, которые образуют совокупность $\{n_1,\ldots,n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$.

К этой задаче сводится один из вариантов проблемы помехоустойчивого распознавания последовательности (1), как структуры, включающей повторяющийся ненулевой вектор, совпадающий с некоторым элементом из заданного алфавита векторов, которая кроме этого вектора содержит неизвестные ненулевые векторы-вставки. В [3] показано, что к решению задачи 1 приводит статистическая формулировка проблемы, если считать, что $\mathbf{e}_n$ в формуле (3) есть выборка из q-мерного нормального распределения с параметрами $(\mathbf{0}, \sigma^2 \mathbf{I})$, где $\mathbf{I}$ - единичная матрица, а в качестве критерия решения задачи использовать максимум функционала правдоподобия.

Редуцированная экстремальная задача

Нетрудно аналитически убедиться, что для любых допустимых $\mathcal{M}_1$ и $\mathcal{M}_2$ при фиксированном векторе $\mathbf{u} \in \mathcal{A}$ минимум функционала (4) по неизвестным векторам $\mathbf{w}_n$, $n \in \mathcal{M}_2$, доставляется векторами $\hat{\mathbf{w}}_n = \mathbf{y}_n$, $n \in \mathcal{M}_2$, и равен

$$S_{\min}(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{N}} \|\mathbf{y}_n\|^2 - \sum_{n \in \mathcal{M}_1} \{2\langle \mathbf{y}_n, \mathbf{u} \rangle - \|\mathbf{u}\|^2\} - \sum_{n \in \mathcal{M}_2} \|\mathbf{y}_n\|^2, \tag{5}$$

где $\langle \cdot, \cdot \rangle$ – скалярное произведение векторов.

Первый член в правой части равенства (5) является константой. Поэтому имеем следующую оптимизационную задачу, к которой сводится минимизация целевой функции (4).

Задача SVVGA (Searching for a Vector in a Given Vector Alphabet). *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, множество (алфавит) $\mathcal{A}$, $|\mathcal{A}| = K$, векторов из $\mathcal{R}^q$ и натуральные числа M_1 и M_2. *Найти*: вектор $\mathbf{u} \in \mathcal{A}$ и непересекающиеся подмножества $\mathcal{M}_1$ и $\mathcal{M}_2$ множества $\mathcal{N} = \{1, 2, \ldots, N\}$, доставляющие максимум целевой функции

$$G(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{M}_1} (2\langle \mathbf{y}_n, \mathbf{u} \rangle - \|\mathbf{u}\|^2) + \sum_{n \in \mathcal{M}_2} \|\mathbf{y}_n\|^2, \tag{6}$$

при условии, что имеют место ограничения $|\mathcal{M}_1| = M_1$, $|\mathcal{M}_2| = M_2$ на мощности искомых подмножеств, а элементы этих подмножеств образуют объединенный набор номеров $(n_1, \ldots, n_M)$ размерности $M = M_1 + M_2$, компоненты которого удовлетворяют ограничениям (2).

Алгоритм решения задачи

Положим

$$g_1(n, \mathbf{u}) = 2\langle \mathbf{y}_n, \mathbf{u} \rangle - \|\mathbf{u}\|^2, \; g_2(n) = \|\mathbf{y}_n\|^2, \; \mathbf{u} \in \mathcal{A}, \; n \in \mathcal{N}.$$

Тогда целевую функцию (6) можно переписать в виде

$$G(\mathcal{M}_1, \mathcal{M}_2, \mathbf{u}) = \sum_{n \in \mathcal{M}_1} g_1(n, \mathbf{u}) + \sum_{n \in \mathcal{M}_2} g_2(n). \tag{7}$$

В настоящей работе показано, что максимум $G_{\max}$ целевой функции (7) вычисляется по формуле

$$G_{\max} = \max_{\mathbf{u} \in \mathcal{A}} G_{\max}(\mathbf{u}), \tag{8}$$

где

$$G_{\max}(\mathbf{u}) = \max_{\mathcal{M}_1, \mathcal{M}_2} G(\mathcal{M}_1, \mathcal{M}_2 \mid \mathbf{u}), \; \mathbf{u} \in \mathcal{A},$$

– условный максимум функции (7), который находится по правилу

$$G_{\max}(\mathbf{u}) = \max_{n \in \omega_M(M)} \max_{t \in \{M_1, M_1+1, \ldots, M\}} G_n(M_1, t, M \mid \mathbf{u}), \tag{9}$$

где $M = M_1 + M_2$. Значения функции $G_n(M_1, t, M \mid \mathbf{u})$, $n \in \omega_M(M)$, $t \in \{M_1, M_1+1, \ldots, M\}$, при каждом фиксированном $\mathbf{u} \in \mathcal{A}$ вычисляются по рекуррентным формулам

$$G_n(l,t,m\,|\,\mathbf{u}) = \begin{cases} g_1(n,\mathbf{u}), & \text{если } l=1,\ t=1,\ m=1, \\ g_1(n,\mathbf{u}) + \max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } l=1,\ t=2,\ldots,M,\ m=t, \\ g_1(n,\mathbf{u}) + \max\limits_{j\in\gamma^-_{m-1}(n)} \max\limits_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}), & \\ & \text{если } l=2,\ldots,M_1, t=l,\ldots,M,\ m=t, \\ g_2(n) + \max\limits_{j\in\gamma^-_{m-1}(n)} G_j(l,t,m-1\,|\,\mathbf{u}), & \\ & \text{если } l=1,\ldots,M_1,\ t=l,\ldots,M,\ m=t+1,\ldots,M, \end{cases}$$

где

$$G_n(0,0,m) = \begin{cases} g_2(n), & \text{если } m=1, \\ g_2(n) + \max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } m=2,\ldots,M, \end{cases}$$

в которых $n \in \omega_m(M)$, причём

$$\omega_m(M) = \{n \,|\, \ 1+(m-1)T_{\min} \le n \le N-(M-m)T_{\min}\},\ m=1,\ldots,M,$$

$$\gamma^-_{m-1}(n) = \{k \,|\, \ \max\{1+(m-2)T_{\min},\, n-T_{\max}\} \le k \le n-T_{\min}\},\ n\in\omega_m(M),\ m=1,\ldots,M.$$

Так как $\{n_1,\ldots,n_M\} = \mathcal{M}_1 \cup \mathcal{M}_2$, поиск непересекающихся подмножеств $\mathcal{M}_1 = \{n_{\mu_1},\ldots,n_{\mu_{M_1}}\}$ и $\mathcal{M}_2 = \{n_{\nu_1},\ldots,n_{\nu_{M_2}}\}$ множества $\mathcal{N}$ эквивалентен поиску объединённого набора $(n_1,\ldots,n_M)$ и одного из подмножеств $\{\mu_1,\ldots,\mu_{M_1}\}$ или $\{\nu_1,\ldots,\nu_{M_2}\}$ множества $\{1,2,\ldots,M\}$; пусть для определённости искомым является подмножество $\{\mu_1,\ldots,\mu_{M_1}\}$. Для каждого фиксированного $\mathbf{u}\in\mathcal{A}$ определим функции:

$$\tilde{G}_j(l,m\,|\,\mathbf{u}) = \max_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}),$$

$$K_j(l,m\,|\,\mathbf{u}) = \arg\max_{s\in\{l-1,\ldots,m-1\}} G_j(l-1,s,m-1\,|\,\mathbf{u}),$$

$$l=2,\ldots,M_1,\ \ m=l,\ldots,M,\ \ j\in\gamma^-_{m-1}(n),$$

где $n\in\omega_m(M)$;

$$I_n(l,t,m\,|\,\mathbf{u}) = \begin{cases} n, & \text{если } l=1,\ t=1,\ m=1, \\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(0,0,m-1), & \text{если } l=1,\ t=2,\ldots,M,\ m=t, \\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} \tilde{G}_j(l,m\,|\,\mathbf{u}), & \text{если } l=2,\ldots,M_1, t=l,\ldots,M,\ m=t, \\ \arg\max\limits_{j\in\gamma^-_{m-1}(n)} G_j(l,t,m-1\,|\,\mathbf{u}), & \text{если } l=1,\ldots,M_1,\ t=l,\ldots,M,\ m=t+1,\ldots,M, \end{cases}$$

где $n\in\omega_m(M)$;

$$J_n(l,m\,|\,\mathbf{u}) = K_{I_n(l,m,m|\mathbf{u})}(l,m\,|\,\mathbf{u}),\qquad l=2,\ldots,M_1,\ \ m=l,\ldots,M.$$

где $n\in\omega_m(M)$.

Тогда в соответствии с (8) искомый вектор $\hat{\mathbf{u}}$ находится по правилу

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{u}\in\mathcal{A}} G_{\max}(\mathbf{u}) .$$

Последние компоненты оптимальных наборов $(\hat{n}_1,\ldots,\hat{n}_M)$ и $(\hat{\mu}_1,\ldots,\hat{\mu}_{M_1})$ согласно (9) определяются по формуле

$$(\hat{n}_M, \hat{\mu}_{M_1}) = \arg\max_{n\in\omega_M(M)} \max_{t\in\{M_1,M_1+1,\ldots,M\}} G_n(M_1,t,M\,|\,\hat{\mathbf{u}}) .$$

Остальные компоненты (при $M_1 > 1$) искомых наборов находятся по следующему правилу:

$$\hat{n}_{m-1} = \begin{cases} I_{\hat{n}_m}(M_1,\hat{\mu}_{M_1},m\,|\,\hat{\mathbf{u}}), & m = M, M-1,\ldots,\hat{\mu}_{M_1}, \\ I_{\hat{n}_m}(l,\hat{\mu}_l,m\,|\,\hat{\mathbf{u}}), & l = L-1,\ldots,2,\ m = \hat{\mu}_{l+1}-1,\ldots,\hat{\mu}_l; \end{cases}$$

$$\hat{\mu}_{l-1} = J_{\hat{n}_{\hat{\mu}_l}}(l,\hat{\mu}_l\,|\,\hat{\mathbf{u}}), \qquad l = L, L-1,\ldots,2.$$

При этом, если $\hat{\mu}_2 - \hat{\mu}_1 > 1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1,\hat{\mu}_1,m\,|\,\hat{\mathbf{u}}), \qquad m = \hat{\mu}_2-1,\ldots,\hat{\mu}_1+1,$$

а если $\hat{\mu}_1 > 1$, то

$$\hat{n}_{m-1} = I_{\hat{n}_m}(1,m,m\,|\,\hat{\mathbf{u}}), \qquad m = \hat{\mu}_1,\ldots,2.$$

Временная сложность алгоритма решения задачи SVVGA есть величина $O[\min\{M_1,M_2\}K(M_1+M_2)^2(T_{\max}-T_{\min}+q)N]$.

Алгоритм решения задачи SVVGA лежит в основе процедуры помехоустойчивого распознавания структурированных данных в виде векторных последовательностей, включающих квазипериодически повторяющийся ненулевой информационно значимый вектор, совпадающий с некоторым вектором из заданного алфавита векторов евклидова пространства. Эта алгоритм гарантирует оптимальность решения по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин.

Заключение

Рассмотренная задача входит в большое семейство актуальных задач [3-5], к которым сводятся типовые проблемы помехоустойчивого off-line анализа и распознавания структурированных данных в виде числовых и векторных последовательностей, включающих повторяющиеся, чередующиеся и перемежающиеся информационно значимые фрагменты или векторы. В настоящей работе представлено алгоритмическое решение одной из таких ранее неизученных задач: обоснован точный полиномиальный алгоритм, который является ядром помехоустойчивого алгоритма распознавания.

Благодарности

Работа поддержана грантами РФФИ 09-01-00032, 07-07-00022 и грантом АВЦП Рособразования 2.1.1/3235.

Литература

[1] Kel'manov A.V., Khamidullin S.A. Recognizing a Quasiperiodic Sequence Composed of a Given Number of Identical Subsequences // Pattern Recognition and Image Analysis, 2000. Vol.10, No.1. P. 127-142.

[2] Kel'manov A.V., Jeon B. A Posteriori Joint Detection and Discrimination of Pulses in a Quasiperiodic Pulse Train // IEEE Transactions on Signal Processing, Vol. 52, No. 3, March 2004, P. 1-12.

[3] Кельманов А.В. Полиномиально разрешимые и NP-трудные варианты задачи оптимального обнаружения в числовой последовательности повторяющегося фрагмента // Материалы Росс. конф. «Дискретная оптимизация и исследование операций» (Владивосток, 7-14 сентября 2007). – Новосибирск: Изд-во Института математики СО РАН, 2007.– http://math.nsc.ru/ conference/ door07/ DOOR_abstracts.pdf. С. 46-50.

[4] Кельманов А.В. О некоторых полиномиально разрешимых и NP-трудных задачах анализа и распознавания последовательностей с квазипериодической структурой // Доклады 13-ой Всеросс. конф. «Математические методы распознавания образов». Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007г. М.: МАКС Пресс, – 2007. С. 261-264.

[5] http://math.nsc.ru/~serge/qpsl

Информация об авторах

***Алексей Долгушев** – аспирант, Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия, e-mail: dolgushev@math.nsc.ru*

***Александр Кельманов** – д.ф.-м.н., ведущий научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; Новосибирский государственный университет, ул. Пирогова, 2, Новосибирск, 630090, Россия; e-mail: kelm@math.nsc.ru*