ЗАДАЧИ ПОМЕХОУСТОЙЧИВОГО АНАЛИЗА И РАСПОЗНАВАНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ВКЛЮЧАЮЩИХ ПОВТОРЯЮЩИЕСЯ УПОРЯДОЧЕННЫЕ НАБОРЫ ВЕКТОР–ФРАГМЕНТОВ[1]

Александр Кельманов, Людмила Михайлова, Сергей Хамидуллин

***Аннотация**: Рассматриваются некоторые задачи помехоустойчивого off-line анализа и распознавания числовых и векторных последовательностей, включающих повторяющиеся наборы квазипериодических фрагментов или векторов. Обоснованы эффективные алгоритмы решения этих задач, гарантирующие оптимальность решения по критерию максимального правдоподобия, в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных случайных величин.*

***Ключевые слова**: структурированная последовательность, упорядоченный набор вектор-фрагментов, помехоустойчивое обнаружение и распознавание, 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

Объектом исследования настоящей работы являются проблемы анализа и распознавания структурированных данных – числовых и векторных последовательностей, в составе которых имеются повторяющиеся, чередующиеся и перемежающиеся информационно значимые фрагменты или векторы. Предмет исследования – некоторые варианты проблемы помехоустойчивого off-line анализа и распознавания последовательностей, включающих повторяющиеся упорядоченные наборы вектор-фрагментов в качестве структурных элементов, в предположении, что скрытые в шуме фрагменты или векторы из искомых наборов совпадают с компонентами упорядоченного эталонного набора векторов, принадлежащего заданному конечному множеству (словарю). Цель работы – обоснование алгоритмов решения этих задач.

Рассмотрим две содержательные задачи. Пусть в первой из них источник сообщений передает информацию об активном состоянии некоторого физического объекта в виде эталонного набора импульсов, имеющих одну и ту же известную длительность, но различную форму. Каждому импульсу соответствует некоторое промежуточное активное состояние объекта. Порядок импульсов фиксирован. Пассивному состоянию соответствует отсутствие каких-либо импульсов. На приемную сторону через канал передачи поступает последовательность квазипериодически чередующихся импульсов, искаженная аддитивным шумом. Термин «квазипериодически» означает, что интервал между двумя последовательными импульсами не одинаков, а лишь ограничен сверху и снизу некоторыми константами. Моменты времени появления импульсов в принятой (наблюдаемой) зашумленной последовательности

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

неизвестны. Требуется обнаружить упорядоченные наборы импульсов в наблюдаемой последовательности, т.е. определить моменты времени, в которые объект находился в активном состоянии.

Во второй содержательной задаче предполагается, что на приемную сторону поступает информация от различных физических объектов, число которых конечно. Каждому объекту однозначно соответствует известный уникальный упорядоченный векторный набор, элемент которого – результат измерения каких-либо характеристик этого объекта в промежуточном активном состоянии. Число промежуточных активных состояний у физических объектов не одинаково. В пассивном состоянии все измеряемые характеристики равны нулю. Упорядоченная совокупность промежуточных активных состояний соответствует активному состоянию этого объекта в целом. На приемную сторону поступает искаженная шумом квазипериодическая последовательность результатов измерения характеристик от неизвестного объекта. Требуется определить (распознать), от какого объекта поступила информация.

Ситуации, в которых возникают сформулированные содержательные задачи, характерны, в частности, для электронной разведки, геофизики, гидроакустики, телекоммуникации и других приложений. В обеих задачах возможны два случая, когда число принятых импульсов или число ненулевых векторных наборов в последовательности известно и неизвестно. Эти случаи для двух сформулированных содержательных задач проанализированы в настоящей работе.

Формальная постановка задач

Пусть $\mathbf{x}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, где $\mathcal{N} = \{1,2,\ldots,N\}$, – последовательность векторов евклидова пространства. Допустим, что эта последовательность имеет следующую структуру

$$\mathbf{x}_n = \begin{cases} \mathbf{u}_1, & n \in \mathcal{M}_1, \\ \mathbf{u}_2, & n \in \mathcal{M}_2, \\ \ldots, & \ldots, \\ \mathbf{u}_L, & n \in \mathcal{M}_L, \\ \mathbf{0}, & n \in \mathcal{N} \setminus \bigcup_{j=1}^{L} \mathcal{M}_j, \end{cases} \tag{1}$$

где $\bigcup_{j=1}^{L} \mathcal{M}_j \subseteq \mathcal{N}$, причем $\mathcal{M}_i \cap \mathcal{M}_j = \varnothing$, если $i \neq j$.

Положим $|\mathcal{M}_j| = M_j$, $j = 1,2,\ldots,L$, и $\{n_1,\ldots,n_M\} = \bigcup_{j=1}^{L} \mathcal{M}_j$, где $M = \sum_{j=1}^{L} M_j$. В дополнение к этому допустим, что

$$\mathcal{M}_j = \{n_m \mid m \equiv j (\mathrm{mod}\, L),\ 1 \le m \le M\},\ j = 1,\ldots,L, \tag{2}$$

причем элементы набора $(n_1,\ldots,n_M)$, соответствующие номерам ненулевых векторов в последовательности (1), удовлетворяют ограничениям

$$1 \le T_{\min} \le n_m - n_{m-1} \le T_{\max} \le N-1, \quad m = 2,\ldots,M, \tag{3}$$

где $T_{\min}$ и $T_{\max}$ – натуральные числа.

Ограничения (3) устанавливают допустимый интервал между двумя ближайшими номерами ненулевых векторов в последовательности (1). Эти ограничения можно трактовать как условие квазипериодичности повторов ненулевых векторов в последовательности (1).

Из (1)-(3) видно, что последовательность $\mathbf{x}_n$ включает $\lfloor M/L \rfloor$ полных повторов векторного набора $(\mathbf{u}_1,...,\mathbf{u}_L)$ и, возможно, один неполный набор. Элементы повторяющегося набора $(\mathbf{u}_1,...,\mathbf{u}_L)$ будем интерпретировать как информационно значимые векторы. Доступной для анализа будем считать последовательность

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{e}_n, \quad n \in \mathcal{N}, \tag{4}$$

где $\mathbf{e}_n$ – вектор помехи (ошибки измерения), независимый от вектора $\mathbf{x}_n$. Заметим, что $\mathbf{x}_n = \mathbf{x}_n(n_1,...,n_M,\mathbf{u}_1,...,\mathbf{u}_L)$. Положим

$$S(n_1,...,n_M,\mathbf{u}_1,...,\mathbf{u}_L) = \sum_{n\in\mathcal{N}} \| \mathbf{y}_n - \mathbf{x}_n \|^2 \tag{5}$$

и рассмотрим следующие задачи среднеквадратического приближения.

Задача 1a. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4), набор $(\mathbf{u}_1,...,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$ и натуральное число M. *Найти*: набор $(n_1,\ldots,n_M)$ номеров такой, что целевая функция (5) минимальна.

Задача 1b. *Дано*: последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4), набор $(\mathbf{u}_1,...,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$. *Найти*: набор $(n_1,\ldots,n_M)$ номеров и его размерность M такие, что целевая функция (5) минимальна.

Задачи 1a и 1b отражают сущность проблемы оптимального обнаружения по критерию минимума суммы квадратов уклонений заданного повторяющегося набора информационно значимых векторов в ненаблюдаемой последовательности, структура которой описывается формулами (1)-(3). Отличие этих задач состоит в том, что в первой из них число ненулевых информационно значимых векторов считается заданным, а во второй – неизвестным, т.е. является оптимизируемой величиной.

Положим $\mathbf{w} = (\mathbf{u}_1,...,\mathbf{u}_L)$. Допустим в дополнение к (1)-(4), что $\mathbf{w} \in \mathcal{W}$, причем $|\mathcal{W}| = K$, где

$$\mathcal{W} \subset \{(\mathbf{u}_1,...\mathbf{u}_L) \,|\, \mathbf{u}_j \in \mathcal{R}^q, 0 < \|\mathbf{u}_j\|^2 < \infty, j = 1,\ldots,L;\ L \in \{1,...,L_{\max}\}\}. \tag{6}$$

Здесь $\mathcal{W}$ – множество (словарь) векторных наборов (слов) мощности K, размерность которых не превосходит $L_{\max}$.

Рассмотрим еще две задачи среднеквадратического приближения.

Задача 2a. *Дано*: множество $\mathcal{W}$, $|\mathcal{W}| = K$, наборов векторов из $\mathcal{R}^q$, последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4) и (6), а также натуральное число M. *Найти*: векторный набор $\mathbf{w} \in \mathcal{W}$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\ldots,n_M)$.

Задача 2b. *Дано*: множество $\mathcal{W}$, $|\mathcal{W}| = K$, наборов векторов из $\mathcal{R}^q$, последовательность $\mathbf{y}_n \in \mathcal{R}^q$, $n \in \mathcal{N}$, структура которой описывается формулами (1)-(4) и (6), *Найти*: векторный набор $\mathbf{w} \in \mathcal{W}$ такой, что целевая функция (5) минимальна на множестве допустимых наборов $(n_1,\ldots,n_M)$.

Задачи 2a и 2b соответствуют проблеме распознавания последовательностей, включающих повторяющиеся наборы чередующихся векторов, скрытых в ненаблюдаемой последовательности (1).

В задаче 2a число ненулевых векторов в последовательности считается заданным, а в задаче 2b – неизвестным.

Легко установить, что к минимизации функции (5) и к таким же сформулированным выше четырем задачам приводит статистический подход к проблемам обнаружения и распознавания, если считать, что $\mathbf{e}_n$ в формуле (4) 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.

Редуцированные оптимизационные задачи

Раскрывая квадрат нормы в формуле (5), получим

$$S = \sum_{n\in\mathcal{N}} \| \mathbf{y}_n \|^2 + \sum_{j=1}^{L} M_j \| \mathbf{u}_j \|^2 - 2\sum_{j=1}^{L} \sum_{n\in\mathcal{M}_j} \left\langle \mathbf{y}_n, \mathbf{u}_j \right\rangle$$

$$= \sum_{n\in\mathcal{N}} \| \mathbf{y}_n \|^2 + \sum_{m=1}^{M} \| \mathbf{u}_{(m-1) \bmod L+1} \|^2 - 2\sum_{m=1}^{M} \left\langle \mathbf{y}_{n_m}, \mathbf{u}_{(m-1) \bmod L+1} \right\rangle,$$

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

Первое слагаемое в правой части полученного выражения – константа. При фиксированных M и $(\mathbf{u}_1,...,\mathbf{u}_L)$ второе слагаемое также является константой. Поэтому имеем следующие редуцированные оптимизационные задачи, к которым сводятся задачи 1a и 1b.

Задача SRTVS-F (Searching for Recurring Tuples of Vectors in a Sequence, when M is Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$, набор $(\mathbf{u}_1,\ldots,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$ и натуральное число M. *Найти*: набор $(n_1,...,n_M)$ номеров такой, что

$$\sum_{m=1}^{M} \left\langle \mathbf{y}_{n_m}, \mathbf{u}_{l(m,L)} \right\rangle \to \max,$$

где $l(m \mid L) = (m-1) \bmod L + 1$, при ограничениях (3).

Задача SRTVS-NF (Searching for Recurring Tuples of Vectors in a Sequence, when M is Not Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$ и набор $(\mathbf{u}_1,\ldots,\mathbf{u}_L)$ ненулевых векторов из $\mathcal{R}^q$. *Найти*: набор $(n_1,...,n_M)$ номеров такой, что

$$\sum_{m=1}^{M} \{2\left\langle \mathbf{y}_{n_m}, \mathbf{u}_{l(m,L)} \right\rangle - \| \mathbf{u}_{l(m,L)} \|^2\} \to \max, \qquad (7)$$

где $l(m \mid L) = (m-1) \bmod L + 1$, при ограничениях (3).

Точные полиномиальные алгоритмы решения этих редуцированных оптимизационных задач обоснованы в [1-3]. Трудоемкости алгоритмов решения задач SRTVS-F и SRTVS-NF есть величины $O[M(T_{\max} - T_{\min} + q)N]$ и $O[L(T_{\max} - T_{\min} + q)N]$ соответственно.

Задачи 2a и 2в сводятся к решению следующих экстремальных задач.

Задача SVTVP-F (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$, натуральное число M и словарь $\mathcal{W}$, $|\mathcal{W}|=K$, упорядоченных наборов векторов из $\mathcal{R}^q$. *Найти*: векторный набор $\mathbf{w}\in\mathcal{W}$ такой, что выполняется (7), при ограничениях (3).

Задача **SVTVP-NF** (Searching for a Vector Tuple in the Vocabulary of Patterns, when M is Not Fixed). *Дано*: последовательность $\mathbf{y}_0,\ldots,\mathbf{y}_{N-1}$ векторов из $\mathcal{R}^q$ и множество (словарь) $\mathcal{W}$, $|\mathcal{W}|=K$, упорядоченных наборов (слов) векторов из $\mathcal{R}^q$. *Найти*: векторный набор $\mathbf{w}\in\mathcal{W}$ такой, что выполняется (7), при ограничениях (3).

Точные полиномиальные алгоритмы решения этих экстремальных задач обоснованы в [4-5]. Временные сложности алгоритмов решения задач SVTVP-F и SVTVP-NF есть величины $O[KM(T_{\max}-T_{\min}+q)N]$ и $O[KL_{\max}(T_{\max}-T_{\min}+q)N]$ соответственно.

Алгоритмы решения приведенных редуцированных задач лежат в основе алгоритмов помехоустойчивого анализа и распознавания структурированных последовательностей, включающих повторяющиеся наборы чередующихся вектор-фрагментов. Эти алгоритмы гарантируют оптимальность решения как по критерию максимального правдоподобия в случае, когда помеха аддитивна и является гауссовской последовательностью независимых одинаково распределенных величин, так и по критерию минимума суммы квадратов уклонений.

Численное моделирование

Результаты численных экспериментов, представленные ниже в качестве примера, носят чисто иллюстративный характер. Они лишь демонстрируют работу алгоритмов и сущность рассмотренных задач для одномерных последовательностей.

На рис. 1 а изображена сгенерированная последовательность X, включающая 3 повтора набора фрагментов. На рис. 1 б представлена последовательность Y, подлежащая обработке (в этом примере уровень помехи превышает уровень сигнала). На рис. 1 в приведена последовательность $\hat{X}$, полученная с помощью алгоритма обнаружения, в условиях, когда число M задано. Прямоугольными рамками очерчены места расположения обнаруженного набора, найденные алгоритмом в зашумленной последовательности. Числовые данные под графиками соответствуют заданным (рис. 1 а) и найденным (рис. 1 б и 1 в) начальным номерам фрагментов. Рисунок иллюстрирует практически безупречную работу алгоритма в условиях, когда уровень сигнала ниже уровня помехи.

На рис. 2 представлены кривые оценок нормированной среднеквадратической ошибки $e(\sigma)=\mathbf{E}\,\|X-\hat{X}\|^2/e^u$, где $\mathbf{E}$ – символ математического ожидания, e^u – оценка сверху для $\|X-\hat{X}\|^2$. Кривая 1 получена с помощью алгоритма обнаружения при неизвестном числе M фрагментов, а кривая 2 – с помощью алгоритма, ориентированного на ситуацию, когда это число известно. Результаты получены при обработке одних и тех же 25000 сгенерированных последовательностей, в составе которых повторялся набор из трех фрагментов; места расположения фрагментов в последовательностях генерировались с помощью датчика случайных чисел.

Рис. 3 иллюстрирует зависимость от уровня помехи вероятности ошибки распознавания последовательностей, включавших повторы двух различных эталонных наборов, в составе которых имелось по три вектора. Теоретические оценки верхней и нижней границ вероятности ошибки

распознавания $\alpha^u(\sigma)$ и $\alpha^d(\sigma)$ в виде графиков приведены под номерами 1 и 4. Кривые 2 и 3 получены в условиях, когда число M было неизвестно и известно соответственно.

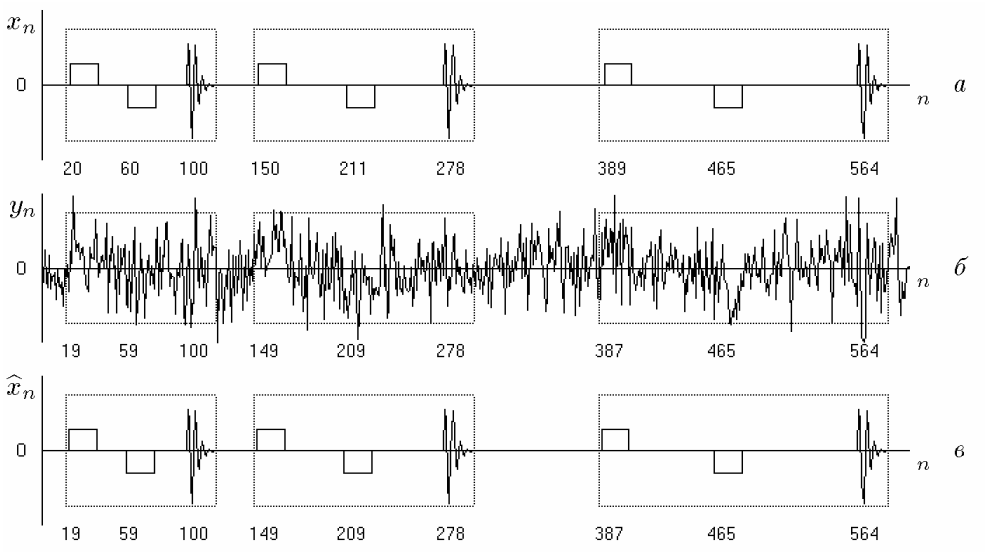

Рис. 1

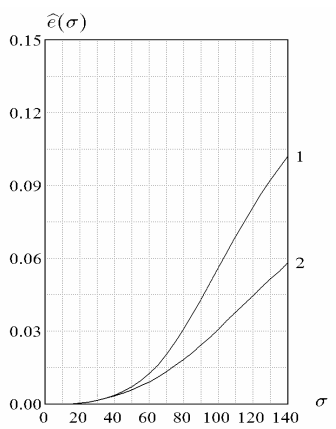

Рис. 2

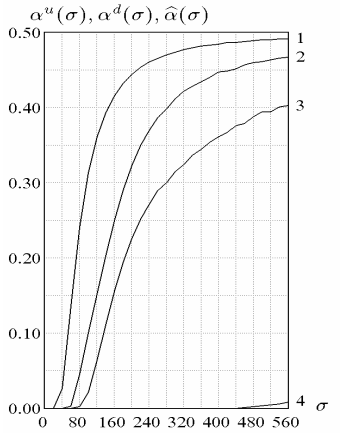

Рис. 3

Оценка вероятности ошибки распознавания при каждом значении σ подсчитана по формуле $\hat{\alpha} = (v_1 + v_2)/2$, где v_1 и v_2 – число неверно опознанных последовательностей, сгенерированных по каждому эталонному набору. Моделировалась байесовская процедура принятия решения с равновероятными гипотезами (наборами). Каждая точка экспериментальной кривой $\hat{\alpha}$ получена в результате усреднения 25000 значений. Рис. 2 и 3 демонстрируют легко доказуемый факт, что ошибка обнаружения и вероятность ошибки распознавания будут меньше в ситуации, когда число ненулевых фрагментов в последовательности известно, чем в ситуации, когда это число неизвестно.

Заключение

Рассмотренные задачи входят в большое семейство актуальных задач [6], 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 векторы или фрагменты. В настоящей работе представлены эффективные алгоритмические решения четырех ранее не изученных задач из этого семейства.

Открытым остается вопрос о разрешимости обобщения рассмотренных задач обнаружения и распознавания на тот случай, когда вместо набора фрагментов, элементы которого упорядочены в соответствии с фиксированным набором векторов, требуется найти набор фрагментов с точностью до всевозможных перестановок элементов фиксированного векторного набора. Алгоритмы решения этих задач представляют значительный интерес для ряда упомянутых во введении приложений. Обоснование алгоритмов решения этих задач представляется делом ближайшей перспективы.

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., Mikhailova L.V., Khamidullin S.A. A Posteriori Detection of a Recurring Tuple of Reference Fragments in a Quasi-Peridic Sequence // Computational Mathematics and Mathematical Physics. 2008, Vol. 48, No. 12, pp. 2276-2288.

[2] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Об одной задаче поиска упорядоченных наборов фрагментов в числовой последовательности // Дискретный анализ и исследование операций. 2009 (принята в печать).

[3] Kel'manov A.V., Mikhailova L.V., Khamidullin S.A. Optimal Detection of a Recurring Tuple of Reference Fragments in a Quasiperiodic Sequence // Numerical Analysis and Applications. 2008. Vol. 1, No.3, pp. 255-268.

[4] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Распознавание квазипериодической последовательности, включающей повторяющийся набор фрагментов // Сибирский журнал индустриальной математики. 2008, Т. 11, №2 (34). С. 74-87.

[5] Кельманов А.В., Михайлова Л.В., Хамидуллин С.А. Алгоритм распознавания квазипериодической последовательности, включающей повторяющийся набор фрагментов // Тез. докл. 15-й международной конф. «Проблемы теоретической кибернетики» (Казань, 2-7 июня 2008). Под ред. Ю.И. Журавлева.- Казань: Отечество, 2008.- С. 45.

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

Информация об авторах

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*

Людмила Михайлова *– к.ф.-м.н., старший научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; e-mail: mikh@math.nsc.ru*

Сергей Хамидуллин *– к.т.н., старший научный сотрудник, Институт математики им. С.Л. Соболева Сибирского отделения РАН, проспект академика Коптюга, 4, Новосибирск, 630090, Россия; e-mail: kham@math.nsc.ru*