Использование параллельных вычислений для построения безызбыточных диагностических тестов

А.Е. Янковская[1,2,3,4], **А.В. Ямшанов**[2]

[1]Томский государственный архитектурно-строительный университет, пл. Соляная 2, Томск, Россия, 634003

[2]Томский государственный университет систем управления и радиоэлектроники, пр. Ленина 40, Томск, Россия, 634050

[3]Национальный исследовательский Томский государственный университет, пр. Ленина 36, Томск, Россия, 634050

[4]Национальный исследовательский Томский политехнический университет, пр. Ленина 30, Томск, Россия, 634050

Аннотация. Использование безызбыточных диагностических методов выявления закономерностей целесообразно для широкого круга проблемных областей, таких как медицина, экобиомедицина, психология, социология, геология, экогеология, строительство, генетика, экономика, обучение и др. Это объясняется продвинутыми инструментами для визуализации результатов распознавания, что особенно важно для вышеупомянутых областей. С другой стороны, построения диагностических тестов имеет высокую вычислительную сложность и теоретически имеет степенной рост сложности от размера исходных данных, причем эта проблема особенно актуальна для отказоустойчивых диагностических тестов. Приводится краткое описание используемых алгоритмов. Описываются теоретические и практические оценки при решении разных задач. Предлагаются различные модели и подходы к параллельной реализации. Обсуждается сравнительный анализ предложенных моделей и подходов.

1. Введение

Впервые принятие решений на основе тестовых методов распознавания образов предложено Ю.И. Журавлевым в публикации [1]. Высокая интерпретируемость результатов, получаемых при использовании тестовых методов распознавания образов привела к их применению для большого количества различных проблеммных областей и обеспечило их развитие сразу несколькими школами по распознаванию образов. В настоящее время, крайне актуальна задача развития тестовых методов распознавания образов для их работы с не надежными исходными данными (с учётом различного рода ошибок в базах данных и знаний, например, ошибок, допускаемых экспертами при создании базы знаний, ошибок введения данных при занесении в базу данных и знаний тех или иных значений признаков и др.). Решение такое задачи целесообразно для широкого круга проблемных областей, где некорректно полученный результат распознавания и принятое в последствие решение может привести к крайне плачевным последствиям, например, медицины, экобиомедицина, психологии, социологии, геологии, экогеологии, строительстве, генетике, экономике и др. Одним из подходов, решающим поставленную задачу, был заложен в работе А.Е. Янковской, посвященной построению отказоустойчивых безызбыточных безусловных диагностических тестов (ОУ ББДТ) [2]. К

сожалению, если применение тестовых методов распознавания образов связано с решением NP-полных задач и их высокой вычислительной сложностью [3] и некоторые исследователи предлагают использование стохастических способов сокращения перебора [4], то примение ОУ ББДТ только усугубляют эту ситуацию, поскольку содержим в своей основе задачи, с которыми не работают уже известные стохастические способы и эвристики для сокращения перебора. Настоящая работа продолжает исследования начатые в публикациях [5, 6] и направлена на исследования возможности применения альтернативных алгоритмов для решения задач, необходимых при построении ББДТ и ОУ ББДТ, например, использующих низкоуровневую оптимизацию программной реализации с учетом современных вычислительных технологий, или сведение задачи в другую форму и решение её методами оптимизации, например, с применением генетических алгоритмов. Ниже рассматривается из наиболее вычислительно сложных задач, необходимая для построения ББДТ и ОУ ББДТ, а именно задача нахождения однократных и h-кратных матричных покрытий [7] и применение параллельных вычислений для её решения.

2. Матричная модель представления данных и знаний. Закономерности в данных и знаниях. Тестовые методы распознавания образов

Поскольку рамки доклада ограничены, раздел содержит только минимально необходимую описание для дальнейшего изложения предлагаемых алгоритмов. За подробным описанием следует обратить к диссертации [8].

Для представления данных и знаний, используемых для построения ББДТ и ОУ ББДТ, применяется матричная модель $M =< Q, R, R' >$, пример матричного представления данных и знаний приведен на рисунке 1, где через X обозначен исследуемый объект.

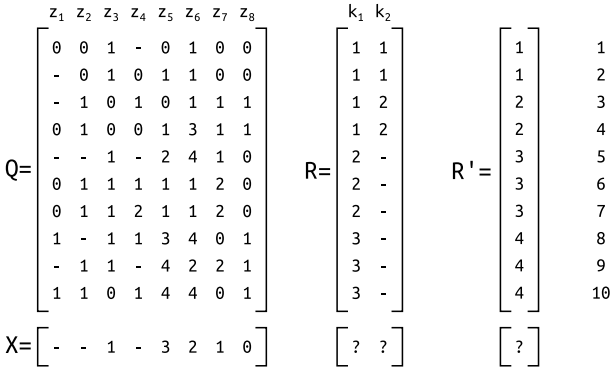


Рисунок 1. Пример матричного представления данных и знаний

Булева или целочисленная **матрица описаний** Q служит для задания описания признаков объектов, строки которой сопоставляются объектам обучающей выборки s_i ($i \in \{1, 2, \ldots, n\}$, где n — количество объектов в обучающей выборке), столбцы — характеристическим признакам z_j ($j \in \{1, 2, \ldots, m\}$, где m — количество характеристических признаков, задающих описание всех объектов обучающей выборки). Если значение характеристического признака несущественно (не влияет на результат распознавания), то этот факт отмечается прочерком ("–") в соответствующем элементе матрицы Q. Кроме того значение характеристического признака может быть задано в интервальном виде $q_{i,j} = [min, max]$.

Строки **матрицы различений** R сопоставляются одноимённым строкам матрицы Q, а столбцы — классификационным признакам k_j ($j \in \{1, 2, \ldots, l\}$, где l — число механизмов классификации), задающим разбиение объектов на классы эквивалентности по каждому

механизму классификации. Элемент $r_{i,j}$ матрицы R задаёт принадлежность i-го объекта одному из выделенных классов по j-му механизму классификации. Для указания факта принадлежности объекта классу используется номер этого класса. С содержательной точки зрения матрицы различений могут нескольких типов, но наиболее часто применяется целочисленная *матрица различений первого типа*, характеризуется вложенностью механизмов классификации, когда каждый последующий столбец задает более подробное разбиение объектов на классы эквивалентности. Строка матрицы различений задаёт **образ** (обобщённый класс). Множество всех неповторяющихся строк матрицы R сопоставлено множеству **выделенных образов**, представленных одностолбцовой матрицей R', элементами которой являются **номера образов**. Матрица R' строится по матрице R путём сопоставления каждой уникальной строки матрицы R номера соответствующего ей образа. Элементами образа являются объекты, представленные строками матрицы описаний, сопоставленными одинаковым строкам матрицы различений R (R').

Если имеется единственный механизм классификации, матрица различений R вырождается в одностолбцовую матрицу R', а образ превращается в класс, что соответствует традиционному представлению знаний в задачах распознавания образов [1].

Под закономерностями в знаниях [8] будем понимать следующие подмножества признаков: константные (принимающие одно и тоже значение для всех образов), устойчивые (константные внутри образа, но не являющиеся константными), неинформативные (не различающие ни одной пары объектов), альтернативные (в смысле включения в ДТ), зависимые (в смысле включения подмножеств различимых пар объектов), несущественные (не входящие ни в один безызбыточный ДТ), обязательные (входящие во все ББДТ), псевдообязательные (входящие в множество используемых при распознавании ББДТ и не являющиеся обязательными), отказоустойчивые (признаки устойчивые к ошибкам измерения) и сигнальные признаки [9] (сигнальные признаки первого и второго рода), а также все минимальные и все (либо часть – при большом признаковом пространстве) безызбыточные различающие подмножества признаков, являющиеся, по сути, соответственно минимальными и ББДТ.

К закономерностям будем относить и ОУ ББДТ, т.е. устойчивые к ошибкам измерения (занесения) значений признаков, описывающих исследуемый объект. Весовые коэффициенты характеристических признаков, также как их информационный вес [1], определяемый на оптимальном подмножестве ОУ ББДТ по признакам, входящим в ОУ ББДТ, относятся к закономерностям.

Для представления сконструированных ББДТ и ОУ ББДТ используется бинарная матрица тестов (T), строки которой сопоставлены тестам, а столбцы — столбцам матрицы Q. Единичное значение в каждой строке матрицы T означает, что соответствующий столбцу признак включён в соответствующий строке тест.

Для выявления закономерностей применяется процедура построения **булевой или целочисленной матрицы импликаций (БМИ)**, задающей различимость объектов из разных образов (классов), столбцы которой сопоставлены столбцам матрицы Q, а строки — всевозможным парам объектов u, w соответственно из разных образов (классов) a, b; $v \in \{1, 2, \ldots, \sigma(Q^a)\}$, $w \in \{1, 2, \ldots, \sigma(Q^b)\}$, где $\sigma(Q^a)$ ($\sigma(Q^b)$) — количество строк в подматрице Q^a (Q^b) матрицы Q.

Далее через $\gamma(R')$ обозначим количество различных элементов в матрице R'. Через $min\ z_i$ и $max\ z_i$ обозначим минимальное и максимальное значение признака z_i соответственно. Строка U_i матрицы U представляет собой значение булевой (целочисленной) **вектор-функции различения**, j-я ($j \in \{1, 2, \ldots, m\}$) компонента $u_{i,j}$ которой вычисляется по формуле: $u_{i,j} = |q^a_{v,j} - q^b_{w,j}|$, где $q^a_{v,j}$ ($q^b_{w,j}$) — значение признака z_j для объекта v (w), а i — определяется по формуле

$$i \in \{1, 2, \ldots, \sum_{r=1}^{\gamma(R')-1} \sum_{s=r+1}^{\gamma(R')} \sigma(Q^r) \cdot \sigma(Q^s)\} \quad (1)$$

Будем говорить, что строка U_a поглощает строку U_b ($U_a \succ U_b$), если и только если

$$(U_a \succ U_b) \leftrightarrow \forall j \in \{1, 2, \ldots, m\}(u_{aj} \geq u_{bj}) \quad (2)$$

Безызбыточной матрицей импликаций назовём такую матрицу U', в которой отсутствуют поглощающие строки.

При построении матрицы U' для вычисления весовых коэффициентов признаков применяется m-компонентный вектор ρ, каждая компонента которого равна сумме значений элементов m-го столбца матрицы U.

Безызбыточной матрицей импликаций (БМИ) назовем такую матрицу U', в которой отсутствуют поглощающие строки.

Процедура построения ББДТ (ОУ ББДТ) включает в себя два этапа:

- Построение матрицы U' по матрицам Q и R (R').
- Поиск безызбыточных (для ОУ ББДТ h-кратных) столбцовых покрытий матрицы U'.

БМИ и фрагмент бинарной матрицы тестов для матриц Q и R приведённых на рисунке 1 представлены на рисунке 2.

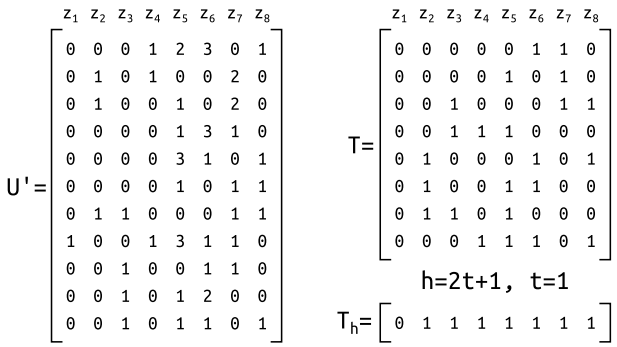


Рисунок 2. Пример БМИ и построенных ББДТ и ОУ ББДТ

Для построения ББДТ (ОУ ББДТ, устойчивых к t ошибкам измерения (занесения) значений признаков в описании исследуемого объекта) необходимо и достаточно, чтобы каждая строка матрицы U' содержала одно (для ОУ ББДТ не менее h ($h = 2t + 1$)) единичных значений. Если это условие не выполняется, то нельзя гарантировать принятие корректного решения (для ОУ ББДТ при наличии t ошибочных значений признаков в описании исследуемого объекта) и требуется расширение признакового пространства (увеличение числа характеристических признаков).

3. Параллельные алгоритмы поиска безызбыточных h-кратных столбцовых покрытий безызбыточной матрицы импликаций

Поиск безызбыточных h-кратных столбцовых покрытий имеет идентичную вычислительную сложность с поиском безызбыточных кратчайших покрытий, но на практике вычислительная сложность поиска безызбыточных h-кратных столбцовых покрытий оказывается намного больше вычислительной сложности поиска безызбыточных кратчайших покрытий, из-за невозможности использования ряда методов, позволяющих сократить число переборов

при построении безызбыточных кратчайших покрытий. Существует два основных подхода к поиску покрытий: в глубину и в ширину. При этом при поиск в ширину достаточно чувствителен к исходным данным, что может быть сглажено применением некоторых шагов от поиска в глубину, что приводит к третьему подходу — гибридному поиску в ширину.

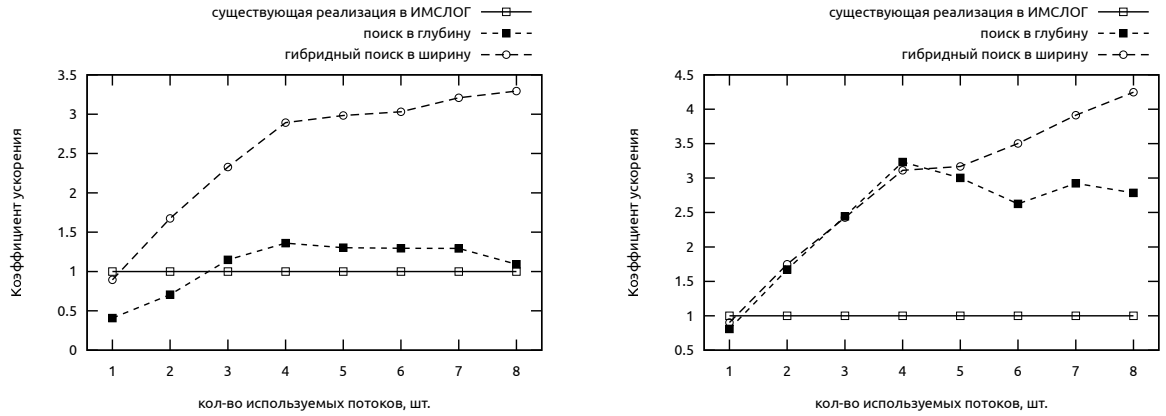
Основная идея поиска в глубину заключается в обходе дерева поиска с одновременным его построением и поиском всех кратчайших неповторяющихся путей, ведущих от корня дерева поиска к листьям. Подобный алгоритм является «жадным», то есть для каждой вершины дерева поиска в первую очередь выбираются столбцы с наибольшим количеством ненулевых значений. Преимуществами поиска в глубину являются: хорошая изученность; большое количество эвристик для поиска однократных покрытий; быстрое получение первого субоптимального решения; возможность сокращения части вычислений путем отсечения менее оптимальных ветвей дерева поиска. К недостаткам следует отнести нетривиальность его распараллеливания и отсутствие подходящих эвристик для поиска h-кратных покрытий, уже существующие эвристики часто сходятся в локальные минимумы, что негативно сказывается на затрачиваемом времени поиска.

Основная идея поиска в ширину заключается в использовании специального генератора блоков битовых последовательностей (масок), последовательно возвращающего все возможные комбинации битовых последовательностей для возрастающего количества признаков и последующую проверку полученных битовых масок на h-кратное покрытие всех строк БМИ. Задача распараллеливания поиска в ширину очевидна и проста, для ускорения вычислений может быть применена технология GPGPU, количество перераспределений памяти минимально, что существенно повышает производительность программной реализации, особенно при использовании параллельных вычислений. При этом недостатками поиска в ширину является: его меньшая изученность; большие временные затраты при поиске однократных покрытий, чем у поиска в глубину; первое решение может быть получено в самом конце выполнения программы.

Большая часть недостатков поиска в ширину сглаживается в гибридном поиске в ширину, где дополнительно выполняется несколько действий, похожих на первый шаг поиска в глубину: все нулевые столбцы удаляются из матрицы, оставшиеся столбцы отсортировываются по количеству ненулевых значений в них. Такая небольшая модификация поиска в ширину, позволяет получать меньшие временные затраты для почти всех исходных данных, по сравнению как с поиском в глубину, так и с поиском в ширину.

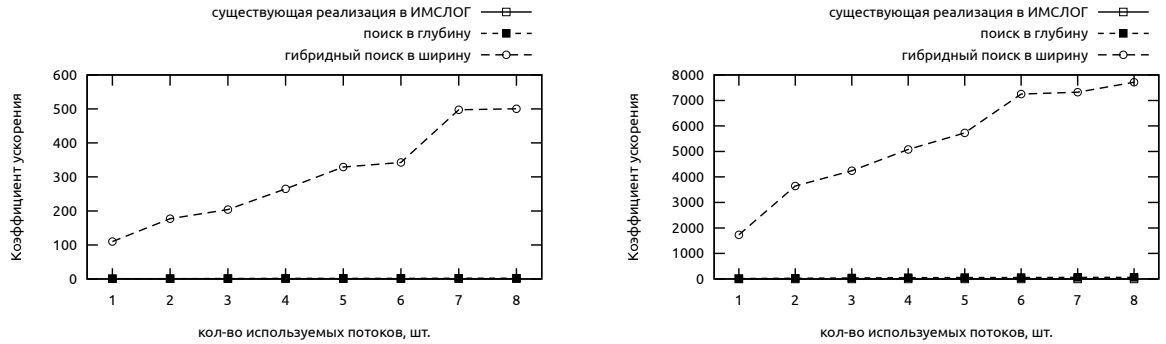
Проведён эксперимент по поиску безызбыточных покрытий с реальными данными из области исследования грунтов (442 строки в БМИ, 22 признака) и дорожно-климатического районирования (82 строки в БМИ, 27 признаков). Для тестирования была использована следующая конфигурация: четырехядерный процессор Intel(R) Core(TM) i7-3770@3.40GHz с Hyper-Threading, 16GiB оперативной памяти DDR3, операционная система Ubuntu Linux 2017.04 kernel 4.10.0, компилятор GCC версии 6.3.0 с флагами компиляции “-std=c++14 -pthread -O2”. Результаты эксперимента представлены на рисунке 3 для области исследования грунтов (а) и дорожно-климатического районирования (б), для сравнения использовался уже существующий алгоритм нахождения покрытий используемый в ИМСЛОГ.

Важным отличием развитых алгоритмов является возможность поиска только n минимальных безызбыточных покрытий, что позволяет отсечь неоптимальные решения на более раннем этапе и приводит к существенному сокращению временных затрат. Результаты соответствующего эксперимента представлены на рисунке 4 для области исследования грунтов (а) и дорожно-климатического районирования (б), для сравнения использовался уже существующий алгоритм нахождения покрытий используемый в ИМСЛОГ.



а) б)

Рисунок 3. Ускорение программы поиска безызбыточных покрытий БМИ в зависимости от количества потоков



а) б)

Рисунок 4. Ускорение программы сокращенного поиска безызбыточных покрытий БМИ в зависимости от количества потоков

4. Заключение

Для решения задачи поиска безызбыточных покрытий были развиты три алгоритма поиска: поиск в глубину, поиск в ширину и гибридный поиск в ширину. Кроме того, для каждого алгоритма было реализованы их параллельные модификации. Сокращение времени выполнения программы уменьшается соизмеримо с увеличением количества используемых ядер и уменьшается более чем в 3 раза на 4-х ядерном процессоре с Hyper-Threading. Предложена полезная модификация алгоритмов отсечение p первых безызбыточных покрытий наименьшего размера, что позволяет сократить время работы программы для некоторых данных на несколько порядков. Результаты экспериментов показали целесообразность применения параллельного алгоритма гибридного поиска в ширину для поиска безызбыточных h-кратных для данных большой размерности (размер матрицы Q более 10^5 строк и 10^2 столбцов) с отсечением p первых безызбыточных покрытий.

5. Благодарности

Работа выполнена при финансовой поддержке РФФИ, проект №16-07-00859а.

6. Литература

[1] Журавлёв, Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики: Выпуск 33. — М.: Наука, 1978. — С. 5-68.
[2] Янковская, А.Е. Выбор оптимального количества признаков при наличии ошибок измерений применительно к задачам медицинской диагностики // Проблемы техники в медицине.
[3] Кудрявцев, В.Б., Андреев А.Е. Тестовое распознавание // Фундаментальная и прикладная математика. - 2009. - Т. 15, №4. - С. 67-99.
[4] Дюкова, Е.В., Прокофьев П.А. Об асимптотически оптимальном перечислении неприводимых покрытий булевой матрицы // ПДМ. 2014. — № 1. — С. 96–105.
[5] Yankovskaya, A. E., Kitler S. V. Parallel algorithm for constructing k-valued fault-tolerant diagnostic tests in intelligent systems // Pattern Recognition and Image Analysis. — 2012. — Vol. 22(3). — P. 473–482.
[6] Yankovskaya, A.E., Yamshanov A.V. An Accelerated Parallel Algorithm for Constructing the Nonredundant Matrix of Implications during the Construction of Fault-Tolerant Nonredundant Diagnostic Tests // Automatic Documentation and Mathematical Linguistics. -2016. - Vol. 50(6). - P. 223–236.
[7] Агибалов, Г.П. Нахождение оптимальных многократных покрытий множеств // Сиб. физ. - техн. ин-т при Том. ун-те. — Томск, 1966. — Вып. 48. — С. 79–86.
[8] Янковская, А.Е. Логические тесты и средства когнитивной графики // LAP LAMBERT Academic Publishing, — 2011. — С. 92.
[9] Yankovskaya, A.E. New Kinds of Regularities in Knowledge and Algorithms of Their Revealing // 7th Open German/Russian Workshop on Pattern Recognition and Image Understanding. August 20—23, 2007. - Ettlingen, Germany.

Usage of Parallel Computations for Irredundant Diagnostic Tests Construction Task

A.E. Yankovskaya[1,2,3,4], **A.V. Yamshanov**[2]

[1]Tomsk State University of Architecture and Building, Solyanaya sq. 2, Tomsk, Russia, 634003
[2]Tomsk State University of Control Systems and Radioelectronics, pr. Lenina 40, Tomsk, Russia, 634050
[3]National Research Tomsk State University, pr. Lenina 36, Tomsk, Russia, 634050
[4]National Research Tomsk Polytechnic University, pr. Lenina 30, Tomsk, Russia, 634050

Abstract. Usage of irredundant diagnostic tests methods of regularities revealing is reasonable for the various problem areas such as medicine, ecobiomedicine, psychology, sociology, geology, ecogeology, civil engineering, genetic, economic, education etc. It is explained by advanced mechanisms of results visualization and substantiation which are especially important in above-mentioned areas. But from another side, diagnostic tests methods has big computational complexity and theoretical power-law growth with input data size and this issue is especially actual for fault-tolerant diagnostic tests. A brief description of used algorithms is given. Theoretical and practical estimations in different tasks are described. Different models and approaches of parallel implementations are suggested. Comparative analysis of suggested models and approaches are discussed.

Keywords: regularities, decision-making, parallel algorithms, performance estimation.