THE PARALLEL IMPLEMENTATION OF THE MULTIHEURISTICAL APPROACH IN THE NUCLEOTIDE SEQUENCE COMPARICON PROBLEM

B.F. Melnikov, doctor of physical and mathematical sciences, professor,
professor of department «Applied mathematics and computer science»
A.G. Panin, candidate of physical and mathematical sciences,
lecturer of department «Applied mathematics and computer science»
Togliatti state university, Togliatti (Russia)

Keywords: multiheuristical approach; nucleotide sequences comparison; Levenshtein distance; strings alignment; parallel programming.

Annotation: We consider the parallel implementation of the multiheuristical approach to determine the nucleotide sequences comparison problem. This approach allows getting a class of metrics on the set of strings, some of which can produce interesting results when applied to the DNA comparison. Several heuristics and the results of their application comparison, as well as the parallel and the sequential algorithm comparison are considered in this article.

УДК 519.68

ПРИМЕНЕНИЕ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ПОДЗАДАЧ ДЛЯ ВЕРШИННОЙ МИНИМИЗАЦИИ НЕДЕТЕРМИНИРОВАННЫХ КОНЕЧНЫХ АВТОМАТОВ

Е.А. Мельникова, кандидат физико-математических наук,
доцент кафедры «Прикладная математика и информатика»
Тольяттинский государственный университет, Тольятти (Россия)

Ключевые слова: дискретная оптимизация, эвристические алгоритмы, кластеризация ситуаций, минимизация недетерминированных конечных автоматов.

Аннотация: В статье рассматривается применение алгоритмов кластеризации подзадач в специальном мультиэвристическом подходе к задачам дискретной оптимизации. А именно, рассматриваются конкретные варианты применения кластеризации ситуаций в задаче минимизации недетерминированных конечных автоматов.

ВВЕДЕНИЕ

Данную статью можно считать продолжением нескольких предыдущих работ, посвящённых применению кластеризации ситуаций в специальном мультиэвристическом подходе к задачам дискретной оптимизации [1–4]. А именно, рассматриваются конкретные варианты кластеризации ситуаций при применении описанного мультиэвристического подхода в задаче минимизации недетерминированных конечных автоматов (НКА).

Задачу вершинной минимизации недетерминированных конечных автоматов можно свести к одной (но важнейшей) из возникающих в ней вспомогательных подзадач. Для любого НКА можно построить *таблицу соответствия вершин* автоматов, являющихся каноническими для исходного и зеркального автоматов [3]. Тогда задачу минимизации можно сформулировать следующим образом. Задана прямоугольная матрица, заполненная элементами 0 или 1. Некоторую пару подмножеств строк и столбцов назовём блоком, если

- на их пересечениях стоят только 1;
- множество нельзя пополнить ни строкой, ни столбцом, не нарушив первого свойства.

Допустимым решением является множество блоков, покрывающих все элементы 1 заданной матрицы. Стоимость решения – количество содержащихся в нем блоков. Цель – выбрать допустимое решение, содержащее минимальное число блоков.

Структура данной статьи следующая. В разделе 2 описывается возможный подход к сравнению эффективности эвристических алгоритмов. В разделе 3 приведены примеры метрик, которые применяются для кластеризации подзадач. В разделе 4 представлены результаты вычислительных экспериментов по применению эвристик, связанных с кластеризацией подзадач. А в заключении приводятся некоторые направления для дальнейшей работы по данной тематике.

ОБ ОДНОМ ПОДХОДЕ К СРАВНЕНИЮ ЭФФЕКТИВНОСТИ ЭВРИСТИЧЕСКИХ АЛГОРИТМОВ

Этот раздел фактически является продолжением соответствующего раздела статьи [1]. Отметим, что какой-либо законченной, «однозначной» теории по поводу подходов к сравнению эффективности эвристических алгоритмов в настоящее время не существует – и вряд ли подобная теория ожидается в будущем [5].

Описанный в данном разделе подход предполагается применять прежде всего в тех случаях, когда вообще невозможно дать гарантию получения оптимального решения. При этом в рассматриваемых задачах дискретной оптимизации (точнее – в их вариантах, зависящих от конкретных алгоритмов псевдослучайной генерации их входных данных) имеется возможность определить некоторую верхнюю границу оптимального решения – вычисляемую согласно алгоритмам генерации входных данных. Отметим, что эта граница совершенно не связана, например, с т.н. границей Хелда-Карпа [6] (т.е. вычисляется совершенно иным способом, значительно более простым) – однако имеет с ней большую аналогию, т.к. используется для тех же самых целей.

Итак, приведём описание предлагаемого варианта сравнения эффективности разных эвристических алгоритмов, предназначенных для решения одной и той же задачи дискретной оптимизации.

Входами каждого эксперимента являются:

– размерность оптимизационной задачи,

– количество случайно сгенерированных задач,

– относительная величина превышения верхней границы оптимального решения (обычно это – величина порядка 0.1, т.е. 10%).

Для каждого конкретного варианта входных данных определяется время работы, в течение которого текущее псевдооптимальное решение достигает этой границы. Это время, усреднённое для задач некоторой размерности, назовем «единицей времени» для данной размерности входных данных.

Тогда итоговые результаты для каждого варианта входных данных можно представить с помощью такой единицы времени. Рассматриваем следующие интервалы времени: 0.1 единицы времени, 0.3 единицы времени, 1.0, 3.0, 10.0, 30.0 и 100.0 единиц времени. Тем самым, получаем необходимую для дальнейшего анализа величину: отношение стоимости полученного решения к первоначальной оценке (верхней границе оптимального решения).

ДВА ВИДА МЕТРИК ДЛЯ ЗАДАЧ ОПТИМИЗАЦИИ

В данном разделе рассматриваются конкретные варианты метрики, которые можно применять в алгоритмах кластеризации ситуаций [1-4] для решения задач дискретной оптимизации.

Для решения подзадач из списка «по аналогии» (т.е. для применения уже выбранного разрешающего элемента, в нашем случае, в случае задачи минимизации НКА – конкретной выбранного блока) должен быть конкретный алгоритм вычисления расстояния между подзадачами. Для этого можно применить следующие 2 варианта.

• В первом случае метрика на множестве подзадач определяется по фактическому расстоянию между двумя подзадачами, вычисляемому обычно на основе двух матриц, описывающих рассматриваемые подзадачи. Далее будем называть такую метрику «матричной эвристикой».

• Во втором случае метрика на множестве подзадач определяется на основе описания подзадач, которое формируется в процессе работы МВГ и состоит из двух списков, а именно: списка выбранных разделяющих элементов; списка запрещённых для будущего выбора разделяющих элементов.

Рассмотрим примеры матричных метрик для задачи минимизации НКА. Для этого введём общее обозначение. Пусть X и Y – некоторые множества, $n=|X \cap Y|$ – число элементов их пересечения, $N=|X \cup Y|$ – число элементов их объединения. Тогда будем писать $\Omega(X,Y)=1-n/N$.

Рассмотрим эвристику для определения метрики на множестве блоков [7]. В результате практического программирования, включавшего проверку нескольких альтернативных вариантов, для данной задачи наиболее удачной признана следующая. Пусть X_1 и X_2 – множества номеров строк 1-го и 2-го блока, Y_1 и Y_2 – аналогичные множества для столбцов. Тогда в качестве метрики используем значение

$$(\alpha\cdot\Omega(X_1,X_2)+(1-\alpha)\cdot\Omega(Y_1,Y_2)),$$

где α – некоторый коэффициент, настраиваемый с помощью применения генетических алгоритмов.

Теперь на основе введенной метрики на блоках определим метрику на подзадачах. Для этого сначала определим метрику для двух произвольных массивов блоков. Пусть $X=\{x_1,\dots,x_m\}$ – первый массив, $Y=\{y_1,\dots,y_n\}$ – второй. Тогда само значение метрики есть

$$\rho(X,Y)=(\textstyle\sum_{i,j}\Omega(x_i,y_j))/(m\cdot n),$$

где i и j принимают все возможные значения. Далее, пусть теперь X_1 и Y_1 – массивы для первой подзадачи, X_2 и Y_2 – соответствующие массивы для второй. Обозначим

$$A=\rho(X_1,X_2),\ B_1=\rho(X_1,Y_2),\ B_2=\rho(X_2,Y_1),$$

и в качестве расстояния между подзадачами будем использовать значение

$$(1+2\cdot\alpha)\cdot A-\alpha\cdot(B_1+B_2),$$

где α, как и ранее, – некоторый коэффициент, настраиваемый с помощью генетических алгоритмов.

При использовании именно этой метрики для решения подзадач «по аналогии» получаются удачные версии программ. Однако данная метрика требует предварительного вычисления расстояний для всех пар блоков, сгенерированных к данному моменту. Поэтому автором чаще используется другая метрика, в чём-то аналогичная метрике, введённой выше для самих блоков.

Итак, приведём ещё один вариант метрики на множестве подзадач. Пусть P_1 – множество клеток матрицы первой подзадачи, имеющих значение 1; P_2 – то же для второй подзадачи. Тогда в качестве метрики используем значение $\Omega(P_1,P_2)$. Отметим, что при этом сами формулы получаются значительно проще, чем в предыдущем варианте. Однако, если на множестве блоков вычисления метрики проведены заранее, то вычисления, соответствующие второму варианту определения метрики на подзадачах, работают существенно дольше.

НЕКОТОРЫЕ РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Далее в настоящем разделе приведены результаты работы компьютера, решающего задачу минимизации недетерминированных конечных автоматов на основе мультиэвристического подхода. В случае минимизации НКА, под размерностью понимается максимум из числа состояний двух канонических конечных автоматов: для заданного регулярного языка и для его зеркального образа. Для каждого варианта размерности специальным образом выбиралось число случайно сгенерированных блоков[8]. В качестве первоначальной оценки стоимости псевдооптимального решения берем число состояний исходного автомата.

[8] Не будем конкретизировать выбираемое число блоков и другие технические детали. Отметим только, что сами алгоритмы случайной генерации входных данных заслуживают большого отдельного обсуждения (с точки зрения репрезентативности для некоторых конкретных предметных областей)

Опишем смысл данных, представленных на приведенных ниже графиках. Во всех 4 случаях были случайно сгенерированы НКА размерности 80–90, имеющие 300–350 блоков разных размерностей. Графики на рис.1 и рис.2 описывают решения, найденные в течение интервала времени, равного 1 специальной единице времени, а графики на рис.3 и рис.4 – в течение интервала времени, равного 3 таким единицам.. Графики на рис.1 и рис.3 соответствуют «матричной» эвристике для вычисления метрики, графики на рис.2 и рис.4 – «списочной» эвристике. На каждом из 4 графиков отмечено по 40 точек, каждая из которых представляет собой описание решения одной из 40 случайно сгенерированных задач минимизации НКА (частных случаев проблемы). *X*-координата каждой точки – относительный результат работы алгоритма без применения эвристик для решения подзадач «по аналогии», а *y*-координата – относительный результат работы алгоритма с применением одной из эвристик.

Прямые *x=1* и *y=1.* на каждом из графиков соответствует относительным результатам, равным 1, полученным этими алгоритмами. Кроме того, на каждом графике проведена прямая *x=y*. Тот факт, что большинство точек двух нижних графиков находятся под прямой *x=y*, свидетельствует об улучшении результатов работы алгоритмов в результате применения каждой из двух эвристик. Отметим также, что вряд ли возникнет необходимость сравнивать эти две эвристики: по-видимому, в anytime-алгоритмах первую из них есть смысл применять при наличии временных ресурсов – т.е. времени, необходимого для вычисления фактического расстояния между подзадачами.

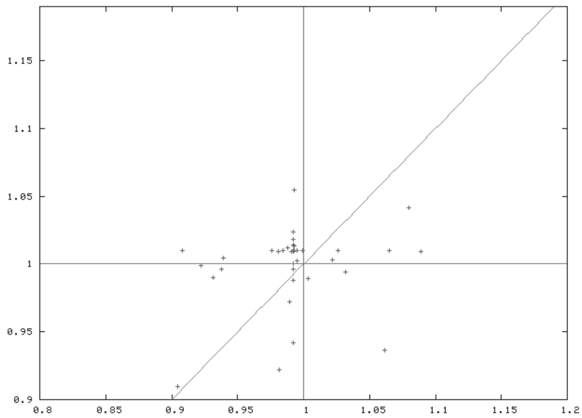

Рис. 1. НКА; матричная метрика; 0.1 единицы времени

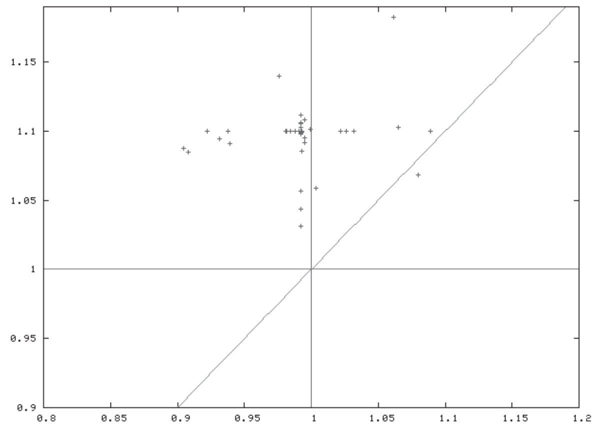

Рис. 2. НКА; списочная метрика; 0.1 единицы времени

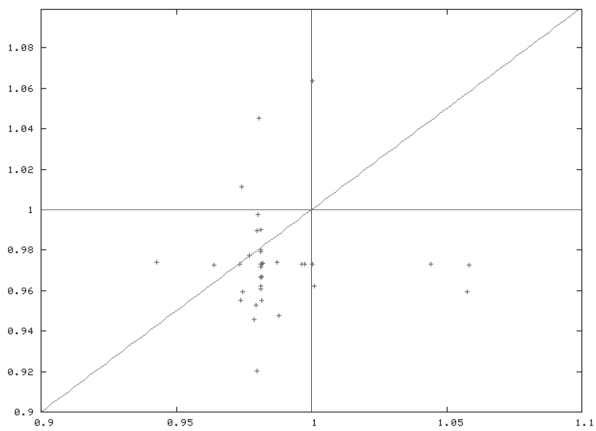

Рис. 3. НКА; матричная метрика; 3 единицы времени

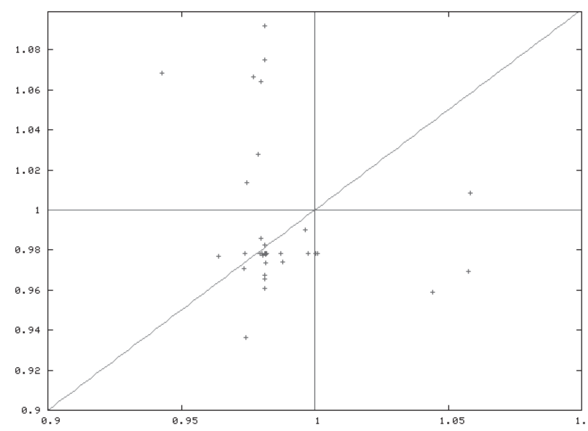

Рис. 4. НКА; списочная метрика; 3 единицы времени

ЗАКЛЮЧЕНИЕ

Одно из основных направлений дальнейшей работы – это разработка и реализация распределенных версий алгоритмов, описанных в данной статье, а также всестороннее исследование репрезентативности входных данных рассматриваемой предметной области и подробное описание подходов к сравнительной оценке качества эвристических алгоритмов.

Работа автора частично поддержана региональным грантом РФФИ № 13-01-97003, а также частично поддержана Федеральной целевой программой «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы (соглашение № 14.В37.21.1934).

СПИСОК ЛИТЕРАТУРЫ

1. Мельников Б. Мультиэвристический подход к задачам дискретной оптимизации. – «Кибернетика и системный анализ» (НАН Украины), 2006, №3, с. 32–42.
2. Melnikov B., Melnikova E., Moseev A., Radionov A. Some specific heuristics for situation clustering problems. – 1st International Conference on Software and

problems. – 1st International Conference on Software and Data Technologies, Setúbal (Portugal) (ICSOFT–2006), Vol.2, p. 272–279.
3. Мельников Б., Мельникова Е. Кластеризация ситуаций в алгоритмах реального времени для задач дискретной оптимизации. – «Системы управления и информационные технологии», 2007, № 2, с.16-19.
4. B.Melnikov, E.Melnikova, A.Moseev Some heuristics for working with search tree in non-deterministic games. – 9th International Workshop on Computer Science and Information Technologies (CSIT'2007). – Ufa, 2008
5. Hromkovič J. Algorithmics for Hard Problems. Introduction to Combinatorial Optimization, Randomization, Approximation, and Heuristics. – Springer, 2003.
6. Held M., Karp R.M. The traveling-salesman problem and minimum spanning trees. Operation Research, 18 (1970) 1138–1162.
7. Melnikov B., Gumayunov V., Radionov A. Some special heuristics for discrete optimization problems. – 8th International Conference on Enterprise Information Systems, Paphos (Cyprus) (ICEIS–2006), pp. 360–364.

APPLYING CLUSTERING ALGORITHMS IN THE PROBLEM OF MINIMIZATION OF NONDETERMINISTIC FINITE AUTOMATONS

E.A.Melnikova, candidate of physical and mathematical sciences, associate professor of the chair «Computer Science and informatics»
Togliatti State University, *Togliatti (Russia)*

Keywords: discrete optimization; heuristic algorithms; clustering situations; minimization of nondeterministic finite automata.

Annotation: In this paper applying situations clustering algorithms in the special multiheuristical approach to discrete optimization problems is considered. Namely, concrete variants of applying situations clustering in the problem of minimization of nondeterministic finite automata are depicted.

УДК 519.688, 536.4.032

К ПОДХОДУ МОДЕЛИРОВАНИЯ ФОРМИРОВАНИЯ НАНОКРИСТАЛЛОВ В ПРОЦЕССЕ КАРБОНИЗАЦИИ ПОРИСТОГО КРЕМНИЯ

Ю.С. Нагорнов, кандидат физико-математических наук, старший научный сотрудник научно-исследовательской части
Б.Ф. Мельников, доктор физико-математических наук, профессор кафедры «Прикладная математика и информатика»
Тольяттинский государственный университет, *г.Тольятти (Россия)*
А.В. Золотов, кандидат физико-математических наук, инженер-программист
ООО «СимбирСофт», г.Ульяновск (Россия)

Ключевые слова: моделирование методом Монте-Карло; рост нанокристаллов; карбид кремния; гетерогенный механизм плавления.

Аннотация: В работе методом Монте-Карло проведено моделирование формирования наночастиц карбида кремния в процессе карбонизации нанокристаллов кремния. Экспериментальные и численные результаты объясняются в рамках термодинамической модели формирования нанокристаллов, в основу которой положен гетерогенный механизм плавления нанокристаллов.