На даний момент є актуальною проблема розробки універсальних і надійних методів і підходів, придатних для обробки інформації різних областей, в тому числі для вирішення проблем, які можуть виникнути в медичній галузі. При лікуванні складних захворювань опорно-рухового апарату, чия етіологія до кінця не розкрита і вимагає додаткового дослідження не є винятком. Для вирішення такого роду завдань має сенс використовувати модифікований алгоритм кластеризації

Ключові слова: кластеризація, модифікація, алгоритм Хамелеон, модифікований метод кластеризації, ієрархія, граф

На данный момент является актуальной проблема разработки универсальных и надежных методов и подходов, пригодных для обработки информации из различных областей, в том числе для решения проблем, которые могут возникнуть в медицинской области. При лечении сложных заболеваний опорно-двигательного аппарата, чья этиология до конца не раскрыта и требует дополнительного исследования, не является исключением. Для решения такого рода задач имеет смысл использовать модифицированный алгоритм кластеризации

Ключевые слова: кластеризация, модификация, модифицированный метод кластеризации (алгоритм Хамелеон), иерархия, граф

УДК 519.7:007.004

АНАЛИЗ ДАННЫХ СЛОЖНЫХ ОБЪЕКТОВ С ПОМОЩЬЮ МОДИФИЦИРОВАННОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ

Т. Б. Шатовская
Кандидат технических наук, доцент*
E-mail: shatovska@gmail.com
О. О. Дорожко*
*Кафедра программной инженерии
Харьковский национальный
университет радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166

1. Введение

Опорно-двигательный аппарат на сегодняшний день досконально не изучен, а соответственно определенный алгоритм оптимального решения наиболее нестабильных и нестандартных течений редких и сложных заболеваний отсутствует.

Попытки систематизировать критерии прогнозирования хода сложных, нестабильных заболеваний, с нестандартной ремиссией изучаются многими исследователями.

В данной области необходима автоматизированная экспертная система прогнозирования хода того или иного заболевания на основании системного подхода к оценке разных показателей. Существующие методы не позволяют принять во внимание всю специфичность данной области, характеризующуюся нелинейными зависимостями, небольшим объемом выборок для анализа, разноразмерными данными, часто непараметрическими величинами. Следует оценить возможность использования модифицированного алгоритма кластеризации в задаче опорно-двигательного аппарата.

Существует множество различных методов, которые могут быть применены для решения поставленной задачи. Однако существует ряд проблем для имеющихся методов:

– проблема обоснования качества результатов анализа. Для различных выборок и данных различные методы оценивания результата могут давать лучший результат;

– во многих областях, а особенно в медицине, имеющиеся данные зашумлены;

– проблема анализа большого числа разнотипных факторов;

– нелинейность взаимосвязей; наличие пропусков, погрешностей измерения переменных.

Так как для различных наборов данных различные методы показывают наилучшие результаты, для каждого отдельного набора данных необходим некий критерий выбора наилучшего метода.

2. Актуальность исследования алгоритмов кластеризации

На сегодняшний день, актуальный алгоритм оптимального решения для исследований это:

– необходимость организовать на единых принципах и синхронизировать выбор метода кластеризации на основании данных анализируемой выборки;

– потребность унифицировать технологии кластеризации и за счет этого сократить время на выбор метода;

– необходимость обеспечения пользователей качественным решением задачи анализа при различных исследуемых данных;

– постоянное увеличение объема поступающей информации и разнородность этой информации требует развития технологий анализа этих данных;

– отдельные методы кластеризации хорошо работают на соответствующих выборках, но не являются универсальными;

– необходимость анализа сложных выборок, с пересекающимися и накладывающимися классами.

Обзор существующих литературных данных, таких как [1 – 4] показал, что для решения поставленной задачи, а именно лечения сложных заболеваний опорно-двигательного аппарата, необходим некоторый систематический подход, позволяющий выявить заболевание у пациента при нестандартном течении. Соответственно для решения данной проблемы необходим алгоритм, позволяющий выходить из положения применяя различные методы в процессе своей работы.

В последнее время разрабатываются и совершенствуются алгоритмы кластеризации способные обрабатывать большие наборы данных. В них основное внимание уделяется масштабируемости. Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами. К наиболее актуальным алгоритмам относятся: BIRCH, CURE, CHAMELEON, ROCK [5]. Сравнение данных методов представлено в табл. 1 [6].

Таблица 1

Сравнение лучших алгоритмов кластеризации для больших объемов данных

Название алгоритма	Большие объемы данных	Устойчивость к шуму	Масштабируемость	Сложность	Определяет количество кластеров	Кластеры произвольного размера и плотности
BIRCH	+	+	+	$O(n \log n)$	+	-
CURE	+	+-	-	$O(n^2 \log n)$--$O(n^2)$	-	+-
CHAMELEON	+	+-	+	$O(nm + n \log n + n^2 \log m)$	+	+
ROCK	+	-	+	$O(\max(n^2 m_a, n^2 \log n))$	-	-

На основании выполненного анализа существующего состояния развития методов и алгоритмов кластеризации можно сделать следующие выводы:

– реальные данные сильно отличаются по характеристикам исследуемой выборки, следовательно, для оптимального анализа необходимо обрабатывать различные выборки различными методами;

– подход, основанный на выборе метода обработки выборки в соответствии с характеристиками исследуемой выборки позволит сократить время и сложность обработки;

– математическая модель выбора метода кластеризации позволит сократить время на выбор метода [7];

– необходимо проводить анализ сложных выборок, с пересекающимися и накладывающимися классами.

Оптимальным алгоритмом для решения поставленной задачи является модифицированный алгоритм кластеризации Хамелеон.

3. Анализ и описание основных этапов модифицированного алгоритма Хамелеон (модифицированного алгоритма кластеризации)

Проблема разделения графа – это разделение вершин этого графа на p примерно равных частей таким образом, чтобы количество ребер между вершинами из разных классов было минимальным. Данная задача применяется во многих различных областях, включая параллельные научные вычисления или планирование задач. Проблема разделения является NP-полной. Тем не менее большое количество разработанных алгоритмов находят достаточно хорошие разделения. Задача k-way разделения чаще всего решается методом рекурсивной бисекции [8 – 10]. В последнее время появился высокоэффективный метод для k-way разделения графа – многоуровневая рекурсивная бисекция (multilevel recursive bisection (MLRB)). Основная структура многоуровневой рекурсивной бисекции очень простая. В начале граф G огрубляется до нескольких сотен вершин, далее выполняется разделение пополам полученного уменьшенного графа, а затем это разделение проецируется обратно на исходный граф через периодическое восстановление разделения [11].

Многоуровневая парадигма также может быть применена для построения k-way разделения прямо на исходном графе [12]. Граф огрубляется последовательно как и в предыдущей схеме. Но теперь огрубленный граф делится сразу на k частей и это k разделение последовательно восстанавливается до исходного графа. Существует ряд преимуществ выполнения сразу k разделения (чем выполнения последовательного посредством многоуровневой рекурсивной бисекции). Во-первых, огрубление необходимо произвести только единожды, что уменьшает сложность алгоритма и время выполнения. Во-вторых, хорошо известно, что многоуровневая рекурсивная бисекция может работать хуже, чем k-way разделение. Таким образом, метод достижения сразу k-way разделения может выполнить разделения лучше. Следует заметить что сразу рассчитать хорошее k-way разделения тяжелее, чем выполнить хорошую бисекцию. Именно по этой причине наиболее распространенное решение задачи k-way разделения выполняется с помощью рекурсивной бисекции [13].

На стадии огрубления размер графа последовательно уменьшается; на стадии исходного разбиения выполняется k-way разделение уменьшенного графа (6-way в данном случае); на стадии восстановления выполняется проецирование разделения на начальный граф.

Например, простейший метод для вычисления начального разделения в контексте многоуровневого алгоритма – это огрубление графа до k вершин. Тем не менее, на фазе усовершенствования необходимо усовершенствовать k-way разделение, которое значительно более сложное, чем усовершенствование бисекции. Даже для 8-way разделения время выполнения для данной схемы достаточно высоко. Для усовершенствования k-way разделения для $k > 8$ время выполнения становится чрезмерно большим [14].

Хамелеон – это алгоритм динамического моделирования в иерархической кластеризации. Ключевым

моментом в алгоритме Хамелеон является то, что он учитывает одновременно взаимосвязанность и близость при определении одинаковых пар кластеров. Именно это позволяет преодолеть ограничения. Хамелеон использует новый подход при определении степени взаимосвязанности и близости между парами кластеров. При данном подходе алгоритм сам просчитывает внутренние характеристики кластеров, следовательно они не зависят от статических, установленных пользователем моделей и могут автоматически подстраиваться к внутренним характеристикам объединенных кластеров.

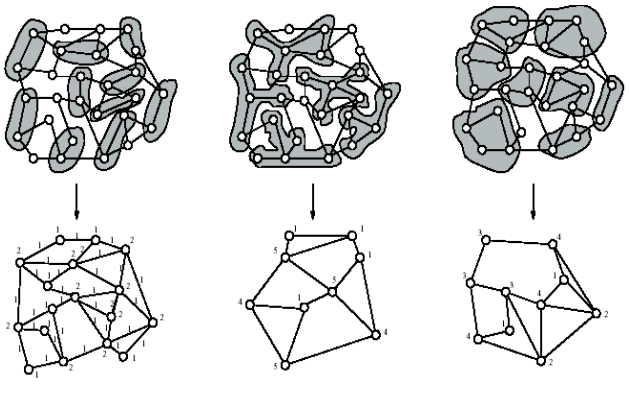
Хамелеон находит кластеры, используя двухфазный алгоритм. На первом шаге Хамелеон использует алгоритм разбиения графа для кластеризации множества на достаточно маленькие подклассы.

На втором шаге используется алгоритм для нахождения естественных кластеров посредством последовательного объединения полученных маленьких подклассов.

Хамелеон представляет объекты посредством часто используемого графа к- ближайших соседей (k-nearest neighbor graph). Это представление данных в виде графа позволяет шкалировать большие объемы данных. Каждая вершина в данном графе представляет один объект данных. Между вершинами существует ребро, если один объект является одним их k ближайших соседей второго объекта. Граф к-ближайших соседей содержит концепцию, что радиус смежности объекта определяется плотностью региона, в котором данный объект находится. Это позволяет выявлять естественные кластеры [1, 7, 15, 16].

Далее строится очередь из последовательно уменьшенных гиперграфов - стадия огрубления (Coarsening Phase). Для огрубления графов может быть применено несколько существующих алгоритмов. На каждом уровне огрубления огрубление заканчивается, как только размер результирующего огрубленного графа уменьшился в 1.7 раз.

В процессе стадии огрубления строится последовательность меньших графов, каждый с меньшим количеством узлов. Огрубление графа может быть достигнуто различными способами. Некоторые возможности показываются на рис. 1 [17].



а б в

Рис. 1. Различные способы огрубления графа:
а — вершинное огрубление (стандартный подход);
б — многовершинное огрубление; в — модифицированное многовершинное огрубление

На третьем шаге выполняется k-way разделение огрубленного графа таким образом, чтобы было удовлетворено ограничение баланса и оптимизирована функция разделения (mincut).

На четвертом шаге выполняется восстановление графа. Разделение огрубленного графа проецируется на следующий уровень исходного графа и выполняется алгоритм улучшения разделения (partitioning refinement algorithm) для улучшения целевой функции, не нарушая ограничение баланса.

Далее определяется показатель схожести между каждой парой кластеров, принимая во внимание их относительную связность и относительную близость. Это позволяет выбрать для объединения кластеры, которые хорошо связаны и достаточно близки. Выбирая кластеры и основываясь на этих двух критериях, Хамелеон преодолевает ограничения существующих алгоритмов, которые оценивают или взаимосвязь или близость [18].

Следовательно, можно выделить следующие стадии:

1. Построение графа. Граф может быть построен симметричный или ассиметричный. Различные виды расстояний могут быть применены при построении графа: Euclidian, Manhattan, Minkowski, SquEuclidian.

2. Огрубление графа (Coarsening). Огрубление графа может быть выполнено следующими методами: Random Matching (RM), Heavy Edge Matching (HEM), Light Edge Matching (LEM) [8, 12, 17, 19].

3. Начальное разделение графа (Initial Partitioning). Существует несколько подходов к разделению графов: графические методы, комбинаторные методы и спектральные методы. Также алгоритмы могут быть выполнены в рамках рекурсивной бисекции, так как большинство методов выполняет деление графа пополам.

4. Восстановление графа (Uncoarsening) и усовершенствование разделения графа (Refinement) Для улучшения разделения графа применяются следующие алгоритмы: Kernighan–Lin (KL), Boundary KL, Fiduccia-Mattheyses (FM), BoundaryFM [20 – 25]. Эти же алгоритмы могут быть применены на этапе разделения, взяв за начальное случайное разделение огрубленного графа.

5. Объединение схожих классов для получения финального разбиения.

4. Исследование завершающего этапа алгоритма – объединение схожих классов для получения финального разбиения

Ключевым шагом является поиск пары подклассов, которые наиболее похожи. На последней итерации определяется показатель схожести между каждой парой кластеров, принимая во внимание их относительную связность и относительную плотность. Это позволяет выбрать для объединения кластеры, которые хорошо связаны и достаточно плотностны [18].

1. Относительная связность и относительная плотность. Этот метод комбинирует относительную связность и относительную плотность, для объединения выбирается пара кластеров, которые максимизируют полученную функцию. Так как целью яв-

ляется объединение классов и с высокой степенью относительной связанности и относительной плотности, логично определить такой функцией их произведение.

Т. е. выбирается такая пара кластеров C_i и C_j, чтобы максимизировать (1):

$$RI(C_i,C_j)*RC(C_i,C_j)^a, \quad (1)$$

где a – определенный пользователем параметр. Если a>1 то большее значение отдается относительной плотности, если a<1 то относительная связность имеет большее значение.

Относительная связность: Обычно в алгоритмах кластеризации оценивается абсолютная связность между кластерами C_i и C_j - сумма весов ребер, соединяющих 2 кластера, которая обозначается как $EC(C_i, C_j)$. Относительная связность между кластерами - это их абсолютная связность, нормированная относительно их внутренней связности. Для получения внутренней связности суммируются ребра, пересекающие разделитель, который разделяет кластер на две примерно равные части.

$$RI(C_i,C_j)=\frac{\left|EC_{\{C_i,C_j\}}\right|}{\frac{1}{2}\left(\left|EC_{C_i}\right|+\left|EC_{C_j}\right|\right)}, \quad (2)$$

Относительная плотность. Абсолютная плотность кластера – это средний вес ребер, соединяющих вершины C_i и C_j. Для получения относительной плотности для пары кластеров абсолютная плотность нормализуется с учетом плотности двух кластеров (3):

$$RC(C_i,C_j)=\frac{\bar{S}_{EC_{\{c_i,c_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}}+\frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}}, \quad (3)$$

где $\bar{S}_{EC_{C_i}}$ и $\bar{S}_{EC_{Cj}}$ средние веса ребер, которые принадлежат к кластерам C_i и C_j и $\bar{S}_{EC_{\{c_i,c_j\}}}$ средний вес ребер соединяющих C_i и C_j.

Обозначения $|C_i|$ и $|C_j|$ количества вершин в каждом кластере.

Данное выражение также нормализует абсолютные плотности для двух кластеров средней взвешенной внутренней плотностью C_i и C_j [18].

2. *Схожесть кластеров(Cluster Similarity).* CS (Cluster Similarity) вычисляется по следующей формуле (4):

$$CS=\frac{|c_{ij}|}{\min(|c_i|,|c_j|)}*\left(\frac{s_{ij}}{\frac{|c_i|}{|c_i|+|c_j|}s_i+\frac{|c_j|}{|c_i|+|c_j|}s_j}\right)^{\alpha}*\left(\frac{\min(s_i,s_j)}{\max(s_i,s_j)}\right)^{\beta}, \quad (4)$$

где $|c_{ij}|$ – количество ребер, которые соединяют вершины подкласса i и вершины подкласса j;

$|c_i|,|c_j|$ – количество ребер внутри подклассу i и j соответственно;

$|s_{ij}|$ – средняя длина ребер, которые соединяют вершины подкласса i и вершины подкласса j;

$|s_i|,|s_j|$ – средняя длина ребер внутри подкласса i и j соответственно;

α,β – задаются пользователем.

Первая часть формулы – это количество ребер, которые соединяют два класса по отношению к количеству ребер в меньшем классе. Это позволяет учесть связанность подграфов. Вторая часть показывает, насколько схожи два подкласса. На каждом шаге объединяется та пара подклассов, в которой данная мера максимальна. Процесс объединения заканчивается, когда количество классов равняется заданной пользователем.

7. Выводы

В данной работе проведен анализ существующих лучших алгоритмов кластеризации, осуществлено сравнение данных алгоритмов и выбран наиболее оптимальный алгоритм кластеризации сложных объектов. Таковым является модифицированный алгоритм динамической кластеризации Хамелеон (модифицированный алгоритм кластеризации). Данный алгоритм состоит из следующих шагов: построение графа, огрубление графа, разделение графа, восстановление и улучшение графа. На каждый из данных этапов предлагаются и реализовываются наборы методов, которые позволяют улучшить качество кластеризации для каждого предложенного конкретного набора данных. Для исследования заболеваний опорно-двигательного аппарата, как сложного объекта, наиболее подходящим методом кластеризации является модифицированный алгоритм Хамелеон (модифицированный алгоритм кластеризации). Проведено исследование завершающего этапа алгоритма - объединение схожих классов для получения финального разбиения данного алгоритма. В дальнейшей работе планируется исследование качества кластеризации сложного объекта - опорно-двигательной системы в экспериментах.

Литература

1. Ляховец, А. В. Исследование результатов применения модифицированного алгоритма хамелеон в области лечения поясничного стеноза. [Текст] / А. В. Ляховец // Восточно-Европейский Журнал передовых технологий. – 2012. – Т. 3, № 11 (57). – С. 13–16.
2. Geisser, Michael E. Spinal canal size and clinical symptoms among persons diagnosed with lumbar spinal stenosis [Text] / Michael E. Geisser, Andrew J. Haig; Henry C. Tong, Karen S. J. Yamakawa, Douglas J. Quint, Julian T. Hoff, Jennifer A. Miner, Vaishali V. Phalke // The Clinical journal of pain. – 2007. – № 23(9). – P. 780–785.

3. Tomkins-Lane Predictors of walking performance and walking capacity in people with lumbar spinal stenosis, low back pain, and asymptomatic controls [Text] / Tomkins-Lane, C. Christy Sara Christensen Holz, Karen S. J. Yamakawa, Vaishali V. Phalke, Doug J. Quint, Jennifer Miner, Andrew J. Haig // Archives of Physical Medicine and Rehabilitation. – 2012. – № 93(4). – P. 647–653.
4. Красиленко, О. П. Лікування синдрому нейрогенної інтермітуючої кульгавості, обумовленого стенозом поперекового відділу хребтового каналу [Текст] / О. П. Красиленко, Ю. Є. Педаченко // Міжнародний неврологічний журнал. – 2011. – № 3. – С. 21–26.
5. Han, J. Data Mining: Concepts and Techniques Second Edition [Text] / J. Han, M. Kamber. – MORGAN KAUFMANN PUBLISHERS, San Francisco, CA, USA, 2006. – P. 354–363
6. Jain, Anil Algorithms for clustering data [Text] / Anil Jain, K. Dubes, C. Richard. – Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. – 304 p.
7. Шатовская, Т. Б. Модификация алгоритма построения графа в алгоритме Хамелеон [Текст] / Т. Б. Шатовская, А. В. Ляховец, И. В. Каменева // Искусств. интеллект. – 2012. – № 3. – С. 480–486.
8. Chan, T. Multilevel generalized force-directed method for circuit placement [Text] / T. Chan, J. Cong, K. Sze. – In Proc. ISPD, ACM New York, NY, USA, 2005 – P. 185–192.
9. Karypis, G. Multilevel graph partitioning schemes [Text] / G. Karypis, V. Kumar. – Minneapolis (Mn): (UMSI research report). Univ. of Minnesota, 1995 – 28 p.
10. Karypis, G. Fast and highly quality multilevel scheme for partitioning irregular graphs [Text] : Intl. Conf. on Parallel Processing / G. Karypis, V. Kumar // SIAM J. Sci. Comput., to appear. – Society for Industrial and Applied Mathematics Philadelphia, PA, USA. – 1998. – Vol. 20, Issue 1. – P. 359–392. – Available at: http://www.cs.umn.edu/_karypis.
11. Бериков, В. С. Современные тенденции в кластерном анализе [Текст] / В. С. Бериков, Г. С. Лбов. – Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. – 26 с.
12. Sumathi, S. Fundamentals of relational database management systems [Text] / S. Sumathi, S. Esakkirajan. – Electronic text data. – Berlin. Heidelberg: Springer-Verlag, 2007. – P. 415–471
13. Thangadurai, Dr. K. A Study On Rough Clustering [Text] / Dr. K. Thangadurai, M. Uma, Dr. M. Punithavalli // Global Journal of Computer Science and Technology. – 2010. – Vol. 10, Issue 5. – P. 55–58.
14. Jain, A. K. Data Clustering: A Review [Text] / A. K. Jain, M. N. Murty, P. J. Flynn // CM Computing Surveys (CSUR). ACM Press, New York. – 1999. – Vol. 31, Issue 3. – P. 255–316.
15. Ляховец, А. В. Экспериментальные результаты исследования качества кластеризации разнообразных наборов данных с помощью модифицированного алгоритма Хамелеон [Текст] / А. В. Ляховец // Вестник запорожского национального университета. – 2011. – № 2. – С. 86–73.
16. Ляховец, А. В. Характеристики выборк данных для выбора k при построении графа к-ближайших соседей [Текст] : VI межд. Науч.-прак. Конф. / А. В. Ляховец // Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій. – Запорожье, 2012. – С. 168–169.
17. Guojun, Gan Data Clustering: Theory, Algorithms, and Applications [Text] / Gan Guojun, Ma. Chaoqun, W. Jianhong. – ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007. – P. 19–320
18. Ляховец, А. В. Исследование динамической кластеризации линейнонеразделимых зашумленных данных с помощью модифицированного алгоритма Хамелеон [Текст] / А. В. Ляховец // Научно-технический журнал “Автоматизированные системы управления и приборы автоматики”. – 2012. – С. 55–62
19. Нейский, И. М. Классификация и сравнение методов кластеризации [Текст] / И. М. Нейский // Интеллектуальные технологии и системы. Сборник учебно-методических работ статей аспирантов и студентов. – 2008 – Вып. 8. – С. 111–122.
20. Salamov, V. Prediction of Protein Secondary Structure by Combining Nearest-neighbor Algorithm and Multiple Sequence Alignments [Text] / V. Salamov, V. Solovyev. – J. Mol. Biol, 1995. – P. 11–15.
21. Якобовский, М. В. Обработка сеточных данных на распределенных вычислительных системах [Текст] / М. В. Якобовский // Вопросы атомной науки и техники. Сер. «Математическое моделирование физических процессов». – 2004. – Вып. 2. – 29 с.
22. Valgaerts, Levi Dynamic load balancing using space-filling curves [Electronic resource] / Technische Universität München, Institut für Informatik. – Levi Valgaerts. – Available at: http://www5.in.tum.de/lehre/seminare/clust_comp/SS05/papers/topic09.doc.
23. Graph Partitioning Algorithms for Distributing Workloads of Parallel Computations Bradford L. Chamberlain Tech. report TR-98-10-03 [Electronic resource / Univ. of Washington, Dept. of Computer Science & Engineering, 1998. –Available at: http://www.cs.washington.edu/homes/brad/cv/pubs/degree/generals.html.
24. Derek, Greene Graph partitioning and spectral clustering” [Electronic resource] / Greene Derek, 2004. – Available at: https://www.cs.tcd.ie/research_groups/mlg/kdp/presentations/Greene_MLG04.ppt.
25. Marks, J. A seed-growth heuristic for graph bisection [Text] / J. Marks, W. Ruml, S. Shieber, J. Ngo; In: Battiti R., Bertossi A. A., editors // Proceedings of Algorithms andExperiments(ALEX98). – Italy: Trento, 1998. – P. 76–87.