УДК 519.7/007/004

А.В. ЛЯХОВЕЦ

ИССЛЕДОВАНИЕ ДИНАМИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ЛИНЕЙНОНЕРАЗДЕЛИМЫХ ЗАШУМЛЕННЫХ ДАННЫХ С ПОМОЩЬЮ МОДИФИЦИРОВАННОГО АЛГОРИТМА ХАМЕЛЕОН

Рассматриваются результаты работы модифицированного алгоритма Хамелеон. Иерархический многоуровневый алгоритм состоит из нескольких этапов: построение графа, огрубление, разделение и восстановление. На каждом этапе используются различные подходы и алгоритмы. Оптимизируется метод выбора k при построении knn графа. Строится математическая модель выбора методов на каждом из этапов алгоритма, основываясь на характеристиках анализируемой выборки.

Введение

Существует много различных методов кластеризации, и каждым из них можно получить различные разбиения исходного множества. Выбор определенного метода зависит от типа желаемого результата. Производительность метода с определенными типами данных зависит от характеристик сервера и технических возможностей программного обеспечения, размера множества. В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие базы 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 [1,2]. Оптимизация и ускорение работы алгоритмов кластеризации является актуальной и востребованной задачей.

Цель данного исследования – разработка математической модели для ускорения и улучшения кластеризации посредством выбора алгоритмов на разных этапах модифицированного алгоритма Хамелеон, соответствующей исследуемым данным. Эта задача явля-

ется актуальной из-за необходимости организации на единых принципах и синхронизации выбора метода кластеризации на основании характеристик анализируемой выборки; потребности унифицировать технологии кластеризации и за счет этого сократить время на выбор метода; необходимости обеспечения пользователей качественным решением задачи анализа при различных исследуемых данных, необходимости анализа сложных выборок с пересекающимися и накладывающимися классами.

1. Модификация алгоритма Хамелеон

Хамелеон – это новый иерархический алгоритм, который преодолевает ограничения существующих алгоритмов кластеризации. Данный алгоритм рассматривает динамическое моделирование в иерархической кластеризации. В нем можно выделить следующие стадии.

Построение графа. В данной работе рассмотрено 2 вида графов: симметричный k-nn граф и асимметричный k-nn граф. При построении графа в модифицированном алгоритме Хамелеон для каждой пары объектов измеряется «расстояние» между ними — степень похожести. Используются следующие меры: евклидово расстояние, квадрат евклидова расстояния, расстояние городских кварталов (манхэттенское расстояние), расстояние Минковского, расстояние Чебышева, степенное расстояние[3,4].

Для решения поставленной задачи построение графа k должно быть выбрано таким образом, чтобы соблюдалось условие его связности. При этом значение k последовательно увеличивается, пока граф не станет связным. Так как данная операция трудоемка и длительна, она нуждается в оптимизации. Для оптимизации выбора k в симметричном и асимметричном графах построены математические модели (рис. 1, а, б).

В качестве управляемых параметров для построения данных моделей зависимости, способных отобразить необходимые характеристики выборки данных, были выбраны количество компонент связности, максимальное расстояние между компонентами связности и количество элементов в компоненте связности. Вторая характеристика вычисляется следующим образом:

$$SetDist = \max\left(\frac{dist(avComponent_i, avComponent_j)}{\max\left(\frac{\max ComponentOstovEdge_{ij}}{ComponentVertexNum_{ij}}\right)}\right), \quad (1)$$

где *avComponent*- центроид компоненты связности; *ComponentOstovEdge* – ребро, соединяющее вершины, которые принадлежат одной компоненте; *ComponentVertexNum* – количество вершин в компоненте.

Данные характеристики не трудоемки в расчете и существует зависимость между ними и значением k.

Математическая модель для оптимизации выбора начального значения k при построении асимметричного k-nn графа:

$$k = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_1^2 + e \cdot x_2^2 + f \cdot x_1 \cdot x_2 + g \cdot x_1^3 + h \cdot x_2^3 + i \cdot x_1 \cdot x_2^2 + j \cdot x_1^2 \cdot x_2, \quad (2)$$

и симметричного графа:

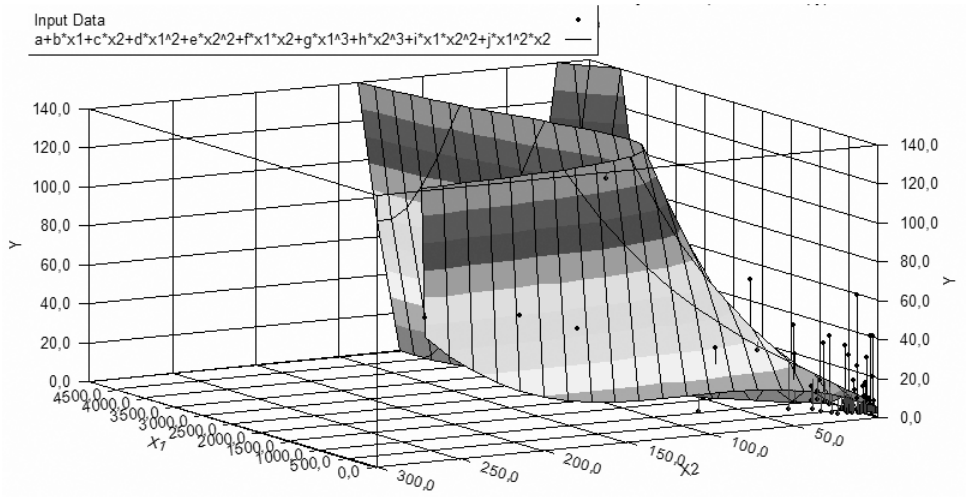
$$k = a + b \cdot x_1 + c \cdot x_1^2 + d \cdot x_1^3 + e \cdot x_2 + f \cdot x_2^2 + g \cdot x_2^3 + h \cdot x_2^4 + i \cdot x_2^5, \quad (3)$$

где x_1 – коэффициент расстояния; x_2 – количество компонент связности. Значения коэффициентов представлены в табл. 1.

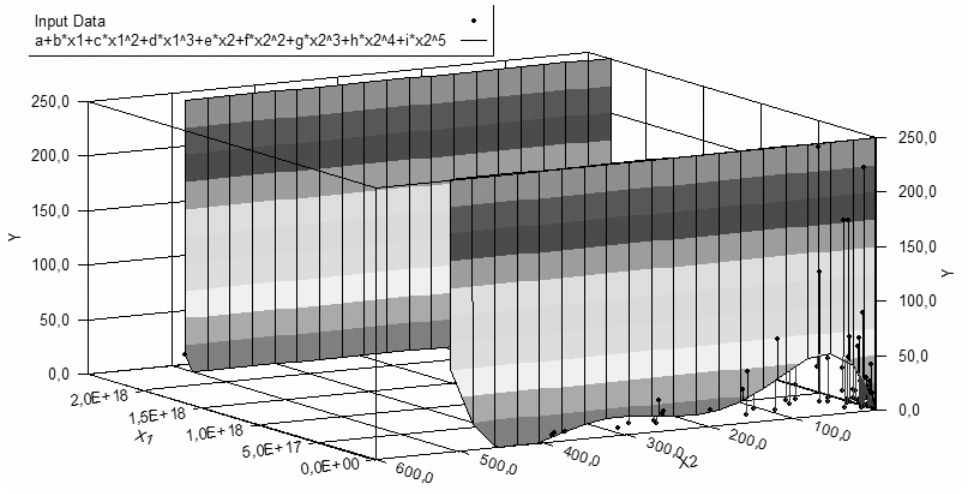
Применение подхода исследовалось на 285 выборках. Применение aknn модели улучшило время выполнения этапа построения графа в 62,45% случаев. В 37,55% случаев время выполнения ухудшилось. Оно ухудшилось лишь в тех случаях, когда k было меньше или равно 3 и время выполнения мало. Следовательно, ухудшение временного показателя несущественно сказывается на производительности метода в целом. Отрицательный результат применения модели получен в 7,71% случаев. В среднем время выполнения улучшилось на 161%. Отрицательным результат считается при получении k существенно больше минимально необходимого для соблюдения условия связности, даже если время построения графа уменьшилось.

Таблица 1

	Aknn граф	Sknn граф
a	4,963024	-0,547360564
b	2,33E-02	-7,46E-14
c	0,42939	1,51E-29
d	-4,45E-05	-6,56E-48
e	-3,86E-03	2,323285358
f	4,18E-04	-3,09E-02
g	1,05E-08	1,55E-04
h	1,14E-05	-3,34E-07
i	1,19E-05	2,61E-10
j	-4,73E-07	



а



б

Рис. 1. Графическое представление описания данных математической моделью

Применение sknn модели улучшило время выполнения этапа построения графа в 69,23% случаев. В 20,51% случаев время выполнения ухудшилось. Отрицательный результат применения модели получен в 5,12% случаев. В среднем время выполнения улучшилось на 169%.

Огрубление графа. В процессе стадии огрубления строится последовательность меньших графов, каждый с меньшим количеством узлов. Огрубление графа может быть достигнуто различными способами [5-8]: случайное паросочетание, паросочетание из тяжелых ребер (HEM), модифицированный алгоритм паросочетания тяжелых ребер (Modified

Heavy Edge Matching - HEM*), паросочетание из наиболее тяжелых ребер (heaviest-edge matching), модифицированное паросочетание из наиболее тяжелых ребер HEM*+, паросочетание легких ребер (Light Edge Matching LEM), паросочетание из тяжелых клик (HCM), сочетане тяжелых треугольников (Heavy-triangle matching HTM), сочетание тяжелых схем (Heaviest Schema Matching HSM), сочетание гиперребер (Hyperedge Coarsening HEC), видоизмененое сочетание гиперребер (Modified Hyperedge Coarsening MHEC), сочетание лучшего (первого) выбора (First Choice Coarsening FCC).

Разделение графа. На данном этапе выполняется разделение огрубленного графа таким образом, чтобы было удовлетворено ограничение баланса и оптимизирована функция разделения (mincut). Разделение может быть выполнено как одновременным разделением на к частей, так и последовательно с помощью рекурсивной бисекции.

Разделение может быть выполнено следующими методами: покоординатное разбиение (Coordinate Nested Dissection (CND)), деление сети с использованием кривых, заполняющих пространство (Space-flling Curve Techniques), алгоритм возрастающего графа (GGP), алгоритм возрастающего графа с учетом выгод (GGGP), уровневое ячеечное разбиение (Levelized Nested Dissection - LND), Seed-Growth bisection, Kernighan-Lin Algorithm (KL), Fiduccia and Mattheyses.

Восстановление графа и улучшение разделения. 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.

Kernighan-Lin Algorithm (KL). KL алгоритм основан на понятии веса - величины, которая определяет выигрыш от перемещения вершины из одного подмножества в другое. Вес рассчитывается для каждой вершины как количество соединений вершины с другим подмножеством, минус количество соединений с подмножеством, в котором вершина находится. Пока есть вершины с положительным весом, алгоритм меняет вершины с максимальным весом местами с вершинами из другого подмножества.

Fiduccia and Mattheyses. По существу схема работы алгоритма такая же, но за один его шаг перемещается только одна вершина, вместо обмена пары вершин, после этого для каждой вершины пересчитывается вес.

Граничный KL и граничный FM (Boundary KL and Boundary FM). Эти алгоритмы в точности повторяют шаги, выполняемые в KL и FM, с тем единственным отличием, что вместо расчета выгоды для всех пар *(u,v)* в графе для KL или всех *u* и всех *v* для FM рассчитываются выгоды только для граничных вершин (т.е. вершин, сопряженных с вершинами из другого класса).

Ключевым шагом является поиск пары подклассов, которые наиболее похожи.

Относительная связность и относительная плотность. Этот метод комбинирует относительную связность и относительную плотность, для объединения выбирается пара кластеров, которые максимизируют полученную функцию.

Схожесть кластеров (Cluster Similarity CS). Данная мера основана на произведении количества ребер, которые соединяют два класса, по отношению к количеству ребер в меньшем классе[9].

2. Описание экспериментальных выборок

Для проверки работоспособности метода необходимо большое количество выборок. Отсутствие реального источника данных требуемого объема, разнообразия и качества вынуждает обратиться к альтернативному источнику. Так как при использовании различных входных данных с определенными статистическими характеристиками производительность и качество кластеризации может сильно отличаться, необходимо проводить анализ на синтетических выборках, созданных специально для данной задачи. Исследований в данной области немного и все они крайне специфичны для рассматриваемых задач.

Существует ряд методов генерации экспериментальных данных, позволяющих провести анализ кластеризации систематически и последовательно. Такие генераторы используют параметризованные модели, которые создают реалистичные данные. Эти генераторы обучены на реальных данных[10].

Helmers и Bunke (2003) разработан для работы с образцами почерка. Baird (2000) и Baird (1993) работали с изображениями. Rogers и др. (2003) работал с 2D изображениями белка. Davidov и соавт. (2004) получали наборы данных, помечая текстовый контент из WWW. Srikant (1999), GSTD (Theodoridis и соавт., 1999) и Jeske и соавт. (2005) также занимались синтетическим созданием данных. GSTD моделирует броуновское движение. Существует ряд скрытых моделей Маркова (Hidden Markov Model -HMM) на основе генераторов данных. Rachkovskij и Kussul (1998) продемонстрировали более общий алгоритм генерации образцов из признаков в пространстве, включая фоновый шум. Pei и Zaiane (2006) занимались получением данных для неконтролируемого обучения и обнаружения выбросов. Van der Walt и Bernard(2007) демонстрируют полезность синтетических генераторов набора данных на основе различных плотностей.

В статье Vineet Chaoji, Mohammad Al Hasan, Saeed Salem и Mohammed J. Zaki «SPARCL: Efficient and Effective Shape-based Clustering» для тестов масштабируемости, а также для создания 3D-данных написан собственный генератор кластеров, основанный на фигурах. Для создания фигуры в 2D случайным образом выбирались точки на канве и добавлялись точки, которые формируют желаемую фигуру. Точкой отсчета для всех фигур была точка (0,0).

Чтобы получить сложные фигуры, использовались фигуры, полученные вращением и смещением (круг, прямоугольник, эллипс, круговые полосы). Генерация 3D фигуры построена на 2D фигуре. Случайным образом выбираются точки для третьей координаты - если координаты *x* и *y* удовлетворяют фигуре, случайным образом выбирается *z*-ось в пределах заданного диапазона.

Такой подход позволяет построить правдоподобные 3D фигуры, а не только несколько слоев 2D фигуры. Как и в случае 2D, комбинируется вращение и смещение 3D фигуры, чтобы получить более сложные фигуры - пример синтетических 3D данных. Как только созданы все фигуры, случайным образом добавляется шум (от 1 до 2%). Показанный на рисунке 3D набор данных имеет 100000 точек и 10 кластеров [11,12].

В данной работе создание 3D фигур выполняется посредством 3D s max studio. Это приложение позволяет сгенерировать трехмерную фигуру необходимой плотности и с необходимым количеством точек. Далее фигура может быть экспортирована. Статистические характеристики полученной выборки будут зависеть от характера фигур, их размера, плотности и расположения. Данные параметры подбираются при создании фигур. Добавление шума в выборку производится непосредственно перед проведением анализа.

Для проведения эксперимента указанным методом было сгенерированно 27 выборок с различными статистическими характеристиками. Выборки и их характеристики представлены в табл. 2. В процессе эксперимента каждая из выборок анализируется в 4 вариантах: без добавления шума, с добавлением 20% шума, 40 и 60% шума.

Также в работе было использовано 47 реальных выборок и 44 выборки, используемые другими авторами при решении задач кластеризации данных.

3. Построение математической модели

Для оптимизации работы модифицированного алгоритма Хамелеон необходимо построить математическую модель зависимости выбора алгоритмов на каждом их этапов модифицированного алгоритма Хамелеон от характеристик обрабатываемой выборки. Математическая модель будет построена на основе исследования 27 экспериментальных выборок и их модификаций посредством добавления шума.

В данной работе представлена модель, в которой используется асимметричный knn граф на этапе построения графа в связи с существенной разницей в трудоемкости по сравнению с симметричным knn графом. В качестве меры схожести при объединении пары подклассов использовалась Cluster Similarity, в качестве меры расстояния - евклидово расстояние.

Для каждой выборки было выполнено сравнение результатов кластеризации с использованием разных алгоритмов на разных этапах модифицированного алгоритма Хамелеон. Лучший результат выбирался на основании таких критериев качества как время выполнения алгоритма и качество кластеризации. Последнее оценивалось на основании следующих

метрик: относительная связность и относительная плотность, Cluster Similarity, SD индекс, PRD индекс, Conn индекс, Silhoette индекс, Dunn индекс.

Таблица 2

№ п/п	Кол-во точек	Maxdist	Кол-во компонент связности	Max матожи-дание	Min матожи-дание	Max диспер-сия	Min диспер-сия	Max разброс	Min разброс
1	266	20,14562	2	0,85	0	127567	22530,6	46,1	17,69
2	2224	30,50736	18	10,03	-0,08	27697655	1651098	49,29	34,09
3	804	161,1586	7	1,78	-0,9	1940398	131320,5	42	17,37
4	1446	197,5728	12	1,16	-1,2	7657729	4567690	52,24	45,78
5	1682	17,14317	39	0,02	-0,01	49026224	9001129	58,06	28,47
6	544	11,71731	32	0,02	-0,03	1791420	295287	46,4	6,47
7	1060	24,27535	28	1,13	-0,17	3032123	12291,01	43,47	5,1
8	682	33,09368	4	6,22	-4,09	649614,3	28661,65	50,51	11,31
9	2650	109,7907	25	0,11	-0,28	25251249	1328711	44,74	12,26
10	511	82,43001	6	0,03	-2,8	1059943	35515,61	72,52	38,78
11	738	211,9365	6	37,22	-0,48	1528520	246586,3	38,97	19,48
12	1250	134,3549	10	0,04	-0,07	2212431	463200,9	48,84	28,27
13	792	8,23338	39	0,09	0	2583547	1627816	69,34	34,57
14	722	0	1	7,08	-0,77	2387575	321328,1	72,31	52,44
15	782	8,956753	49	0	-0,67	6547451	82153,88	144,09	22,27
16	382	0	1	0,7	-0,31	1405572	77981,37	65,17	39,69
17	1292	14,19542	14	2,81	-0,1	10029280	349007,6	34,47	18,81
18	1928	161,9535	29	0,1	-1,01	10586689	4442635	38,43	30,71
19	770	71,53869	10	0,1	-0,21	1355198	46438,56	49,86	13,38
20	1466	218,1389	12	0,26	0	8548895	139608,8	104,86	13,38
21	1751	405,3506	3	0	-0,04	13130396	1260448	86,39	56,84
22	2447	451,3041	61	0	-0,06	26681767	2151437	86,39	59,17
23	84	7,549394	4	1,05	-3,57	20242,3	3634,21	45,11	13,84
24	1196	71,33776	48	73,83	-3,46	1080946	226207,8	29,48	25,78
25	1275	149,7968	34	205,13	-0,09	5886294	1663123	96,49	48,63
26	882	76,0138	5	0,17	-0,06	6779079	125142	83,74	58,52
27	1294	5,692696	107	0,01	0	17868393	9762847	60,01	41,37

В данной работе управляемыми параметрами являются характеристики выборки, такие как количество объектов, вычисляемая характеристика расстояний в выборке, минимальные и максимальные значения матожидания, дисперсии и разброса.

Результирующие значения для выбора алгоритмов на этапах модифицированного алгоритма Хамелеон были закодированы следующим образом: последовательно пронумерованы комбинации алгоритмов, составленные перебором, начиная с полного перебора алгоритмов огрубления, после разделения и восстановления.

В результате была получена следующая математическая модель:

$$k = a + b \cdot x_1 + c \cdot x_1^2 + d \cdot x_2 + e \cdot x_2^2 + f \cdot x_2^3 + g \cdot x_3^2 + h \cdot x_4 + i \cdot x_5 + j \cdot x_5 + \\ + l \cdot x_6 + m \cdot x_6^2 + n \cdot x_7 + n \cdot x_8, \quad (4)$$

где x_1- x_8 соответствуют характеристикам выборок из табл. 2. Значения коэффициентов представлены в табл. 3.

4. Результаты экспериментов

Ключевые аспекты оценивания - это эффективность, надежность, простота и результативность. Расчет времени производился на 1.73 GHz Intel(R) Pentium(R) Dual CPU с 2GB памяти.

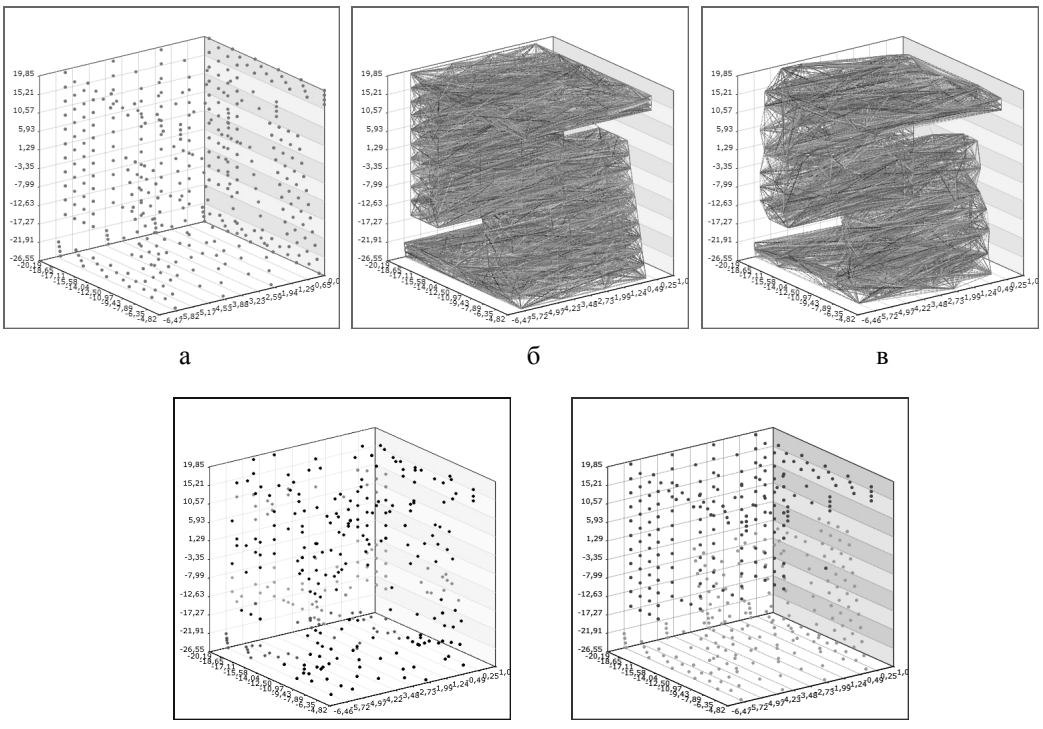
Таблица 3

a	324,765423	i	1,24E-07
b	33,1403	j	1,19E-07
c	0,245459	l	-1,73E-05
d	52,33	m	-1,43E-07
e	43,73	n	1,19E-08
f	4,42E-02	o	-3,75E-06
g	1,62 E-07	i	1,51E-08
h	4, 23	j	-4,77E-08

В данной работе представлена модель, в которой используется асимметричный knn граф на этапе построения графа в связи с существенной разницей в трудоемкости по сравнению с симметричным knn графом. В ходе эксперимента было выявлено, что трудоемкость построения симметричного графа может превышать трудоемкость при построении асимметричного графа в 200 раз.

На основании данных, полученных в ходе эксперимента, можно сделать следующие выводы: алгоритмы этапа огрубления графа не имеют существенного влияния на результат, для больших и не сильно сложных выборок хороший результат от CND, GGP, LND. Для сложных выборок есть смысл использовать FM, несмотря на то, что время выполнения возрастает, особенно на очень больших выборках данных. В этом случае время выполнения можно сократить посредством использования граничного алгоритма для восстановления графа.

Пример обработки одной из выборок представлен на рис 2, 3.



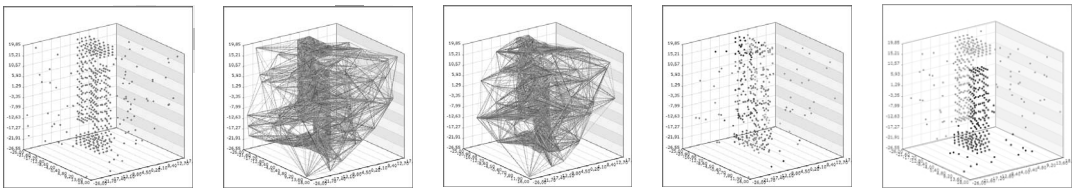
а б в

г д

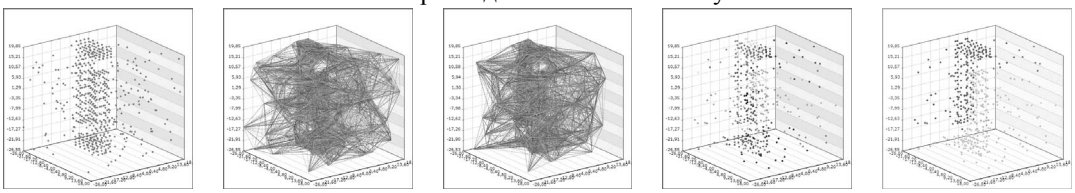
Рис. 2. Результат кластеризации выборки без добавления шума: а – выборка; б – построение графа; в – огрубление графа; г – разделение графа; д – восстановление графа

5. Заключение

В результате исследования на основе содержательного анализа предметной области и существующих решений сформулирована постановка задачи оптимизации процесса кластеризации линейно-неразделимых зашумленных данных.



Выборка с добавлением 20% шума



Выборка с добавлением 40% шума

Рис. 3. Результат кластеризации выборки

Научная новизна. Построено две математические модели зависимости выбора k при построении асимметричного и симметричного графов в модифицированном алгоритме Хамелеон на основании характеристик выборки. Из набора исследуемых характеристик выделены такие: количество компонент связности и рассчитываемая характеристика отношения максимального расстояния между компонентами связности и количеством элементов в компоненте связности. В среднем время выполнения улучшилось на 161 и 169% для асимметричного и симметричного графа соответственно.

Построена математическая модель зависимости выбора алгоритмов на каждом из этапов алгоритма Хамелеон на основании характеристик выборки. Данная модель позволяет сократить время выполнения кластеризации без снижения качества посредством использования алгоритмов, подходящих данной конкретной исследуемой выборке.

На основании проведенного эксперимента можно сделать вывод, что для различных выборок с разными статистическими характеристиками необходимо использовать различные алгоритмы для шагов модифицированного алгоритма Хамелеон. Использование модели особенно критично для больших выборок, где применение многих алгоритмов является дорогостоящим.

Практической значимостью полученных результатов служит подтверждение эффективности подхода при практическом применении результатов исследования на экспериментальных данных и реальных данных в медицинской сфере [13].

В дальнейшем планируется построение математической модели для анализа всех представленных алгоритмов на разных этапах модифицированного алгоритма Хамелеон.

Список литературі: 1. *Чубукова И.А.* Data Mining БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий - ИНТУИТ. 2008 **2.** *Osmar R. ZaË_ane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang* On Data Clustering Analysis: Scalability, Constraints and Validation **3.** *George Karypis, Eui-Hong (Sam) Han, Vipin Kumar*, Chameleon: Hierarchical Clustering Using Dynamic Modeling, Computer. Vol. 32, no. 8. P. 68-75. Aug. 1999, doi:10.1109/2.781637 **4.** *Han J., Kamber M.* Data Mining: Concepts and Techniques Second Edition MORGAN KAUFMANN PUBLISHERS 2006 **5.** *Karypis G. and Kumar V.* Multilevel k-way Partitioning Scheme for Irregular Graphs JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING 48, 96–129 (1998). **6.** *Бувайло Д.П.* Быстрый высокопроизводительный алгоритм для разделения нерегулярных графов // Вісник Запорізького державного університету. 2002. № 2.**7.** *Brian* Read Advances in Databases: 18th British National Conference on Databases, BNCOD 18 Chilton, UK, July 9-11, 2001. Proceedings (Lecture Notes in Computer Science). **8.** *Karypis G. and Kumar V.* Multilevel k-way Partitioning Scheme for Irregular Graphs published electronically. 1999. Society of Industrial and Applied mathematics. **9.** *Ляховец А.В., Лесная Н.С., Шатовская Т.Б* Исследование эфективности динамической кластеризации линейнонеразделимых зашумленных данных // Системы обработки информации. 5(86)’ 2010. С. 86-91. **10.** *Parul Agarwal,M. Afshar Alam,Ranjit Biswas* Issues, Challenges and Tools of Clustering Algorithms IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2. 2011. **11.** *Коршунов Ю.М.* Получение многомерной статистической выборки с заданными корреляционными свойствами ISSN 1995-4565 // Вестник РГРТУ. 2008. Вып. 23. **12.** *Vineet Chaoji, Mohammad Al Hasan, Saeed Salem, Mohammed J. Zaki.* SPARCL: Efficient and Effective Shape-Based Clustering. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy. P. 93-102, IEEE Computer Society, 2008. **13.** *Ляховец А.В.* Исследование результатов применения модифицированного алгоритма хамелеон в области лечения поясничного стеноза // Восточно-европейский журнал передовых технологий. 2012. 3/11(57).

Поступила в редколлегию 21.03.2012

Ляховец Алена Витальевна, мл. научн.сотр. кафедры ПИ ХНУРЭ. Научные интересы: математическое моделирование и анализ данных. Адрес: Украина, 61000, Харьков, пр. Ленина, 12, кв. 58, тел. (066)3256098.