УДК 665.9

Т. Б. ШАТОВСКАЯ, И. В. КАМЕНЕВА

Харьковский Национальный университет радиоэлектроники

ПОСТРОЕНИЕ ГРАФА СВЯЗНОСТИ В АЛГОРИТМЕ КЛАСТЕРИЗАЦИИ СЛОЖНЫХ ОБЪЕКТОВ

Аннотация. В статье представлена модификация алгоритма Хамелеон. Алгоритм Хамелеон состоит из следующих этапов: построение графа, огрубление, разделение и восстановление. На каждом из этапов могут быть использованы различные подходы и алгоритмы. Рассмотрено 2 вида графов: симметричный k-nn граф и ассиметричный k-nn граф.

Ключевые слова: кластеризация, алгоритм Хамелеон, построение графа, связность, k-ближайших соседей.

Анотація. У роботі представлений модифікований алгоритм Хамелеон. Алгоритм Хамелеон побудований з таких етапів: побудова графа, огрубіння, поділ та відновлення. На кожному з цих етапів можуть бути використані різні підходи та алгоритми. Головною метою роботи є дослідження з покращення этапу побудови через оптимізацію алгоритму вибору k під час побудови графа k найближчих сусідів. Розглянуто 2 види графів: симетричний k-nn граф та асиметричний k-nn граф.

Ключові слова: кластеризація, алгоритм Хамелеон, побудова графа, зв’язність, k-найближчих сусідів.

Annotation. In the article, modification of Chameleon algorithm is presented. Chameleon algorithm consists of the following stages: graph construction, coarsening, partitioning and uncoarsening. At each of these steps, different algorithms and approaches can be used. The main goal of this work is investigation and improvement of graph construction stage. This can be done by modification of k-selection algorithm during k-nn graph construction. It is considered two kinds of graphs: symmetric and asymmetric.

Key words: clustering, Chameleon algorithm, graph construction, connectivity, k-nearest neighbors.

Введение

На данный момент весьма активно исследуются различные методы кластеризации. Каждым из целого множества имеющихся методов 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. Модификация алгоритма построения графа в алгоритме Хамелеон

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]. Главной целью работы является исследование и улучшение этапа построения графа посредством оптимизации алгоритма выбора k при построении графа k ближайших соседей.

1. Модифицированный алгоритм 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:

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).
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. Эти же алгоритмы могут быть применены на этапе разделения, взяв за начальное случайное разделение огрубленного графа.
5. Объединение схожих классов для получения финального разбиения.

Целью построения графа является соединение точек локальных соседей. Точки соединяются в зависимости от типа графа.

- Граф эпсилен-окрестности (Epsilon-neighborhood graph). Две вершины графа соединены, если расстояние между рассматриваемыми объектами меньше эпсилен. Данный граф может быть взвешенным и не взвешенным. В случае взвешенного графа вес ребра равняется значению схожести соседних точек (не расстоянию). Параметр данного графа – эпсилен – устанавливается пользователем.
- Полностью связный граф (completely connected graph). Граф может быть получен из графа эпсилен-окрестности установкой эпсилен в максимальное значение.

- Симметричный граф k ближайших соседей (symmetric k-nearest neighbor graph(k-nn)): две вершины x, y соединены, если x находится среди k ближайших соседей y и наоборот.
- Ассиметричный граф k ближайших соседей (mutual k-nearest neighbor graph): две вершины x, y соединены, если x находится среди k ближайших соседей y или наоборот [2].

2. Введение в k-nn граф

Задача графа k ближайших соседей определена следующим образом: дано множество точек P из n точек в R^d и положительное целое число $k \leq n-1$, рассчитать k ближайших соседей для каждой точки P. Более формально задача может быть представлена следующим образом: пусть $P=\{p_1,p_2,\ldots p_n\}$ множество точек в пространстве R^d где $d \leq 3$. Для каждой вершины $p_i \in P$ пусть N^k_i k точек из P ближайших к p_i. Граф k ближайших соседей (k-nearest neighbor graph (k-NNG)) - это граф где множество вершин $\{p_1,p_2,\ldots p_n\}$ и множество ребер $E= \{(p_i, p_j): p_i \in N^k_i$ или $p_j \in N^k_i\}$ [3]. Следует отметить, что это ассиметричный граф k ближайших соседей, так как отношения близости ассиметричны. P_i может быть среди ближайших соседей p_j, но p_j нет. В симметричном графе p_i и p_j будут соединены ребром только в том случае, если каждая из них находится среди k ближайших соседей другой вершины.

В данной работе рассмотрено 2 вида графов: симметричный k-nn граф и ассиметричный k-nn граф. При построении графа для каждой пары объектов измеряется «расстояние» между ними — степень похожести. В данном случае, чем больше сходство меду двумя объектами - тем тяжелее будет ребро между ними.

Еще одним важным параметром при построении графа является k – количество соседей, с которыми будет связанна каждая из вершин. Граф называется *связным*, если в нем для любых двух вершин имеется маршрут, соединяющий эти вершины. При решении поставленной задачи для построения графа k должно быть выбрано таким образом, чтобы соблюдалось условие связности построенного графа. Но слишком большое значение k очень сильно увеличивает вычислительную дороговизну метода и время выполнения не только этапа построения графа, а и всех последующих этапов. Самым простым подходом для выбора k является (2.1), но и данный метод имеет вышеперечисленные недостатки.

$$k = \sqrt{n}$$

На практике применяется два принципиально различных порядка обхода, основанных на поиске в глубину и поиске в ширину соответственно.

Поиск в ширину. Вначале все вершины помечаются как новые. Первой посещается вершина a, она становится единственной открытой вершиной. В дальнейшем каждый очередной шаг начинается с выбора некоторой открытой вершины x. Эта вершина становится активной. Далее исследуются ребра, инцидентные активной вершине. Если такое ребро соединяет вершину x с новой вершиной y, то вершина y посещается и превращается в открытую. Когда все ребра, инцидентные активной вершине, исследованы, она перестает быть активной, и становится закрытой. Если на данном этапе остались незакрытые вершины - то граф несвязный.

Поиск в глубину. Главное отличие от поиска в ширину состоит в том, что при поиске в глубину в качестве активной выбирается та из открытых вершин, которая была посещена последней. Основной алгоритм тот же, что и в случае поиска в ширину, только нужно очередь заменить стеком, а процедуру BFS - процедурой DFS.

Общая оценка трудоемкости для алгоритмов одинаковая - $O(m+n)$.

3. Оптимизация выбора k для построения k-nn графа

Для оптимизации выбора начального параметра k при построении k-nn графа необходимо построить математическую модель зависимости k от характеристик обрабатываемой выборки. Построение математической модели выполнялось на наборе экспериментальных выборок. Набор выборок состоит из 132 выборок, среди них 33 уникальных выборки и 3 вариаций каждой из них полученной путем добавления 20%, 40% и 60% шума. Эксперимент так же проводился на наборах экспериментальных и реальных выборок полученных с ресурсов обмена наборами данных.

Целью данных экспериментов был выбор управляемых параметров данной модели зависимости, способных отобразить необходимые характеристики выборки данных. В рамках работы было проведено 3 эксперимента для выбора управляемых параметров.

- В первом эксперименте анализировались такие характеристики как: количество объектов в выборке, минимальные и максимальные значения математического ожидания, дисперсии и разброса. Зависимости между данными параметрами и значением k не выявлено.
- Во втором эксперименте в качестве управляемого параметра были выбраны длина наибольшего остового ребра полносвязного графа и среднее значение длины всех остальных ребер остова.

Данные характеристики показывают зависимость, но использование данного подхода не является целесообразным в связи с трудоемкостью построения остова полносвязного графа.

- В третьем эксперименте в качестве характеристики использовались количество компонентов связности, максимальное расстояние между компонентами связности и количество элементов в компоненте связности.

В результате исследования была построена математическая модель для оптимизации выбора начального значения k при построении ассиметричного k-nn графа. Модель для ассиметричного k-nn графа имеет следующий вид и представлена на (рис. 1):

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

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

Таблица 1 – Значения коэффициентов модели для определения k в k-nn графе.

α	4,963024	f	4,18E-04
b	2,33E-02	g	1,05E-08
c	0,42939	h	1,14E-05
d	-4,45E-05	i	1,19E-05
e	-3,86E-03	j	-4,73E-07

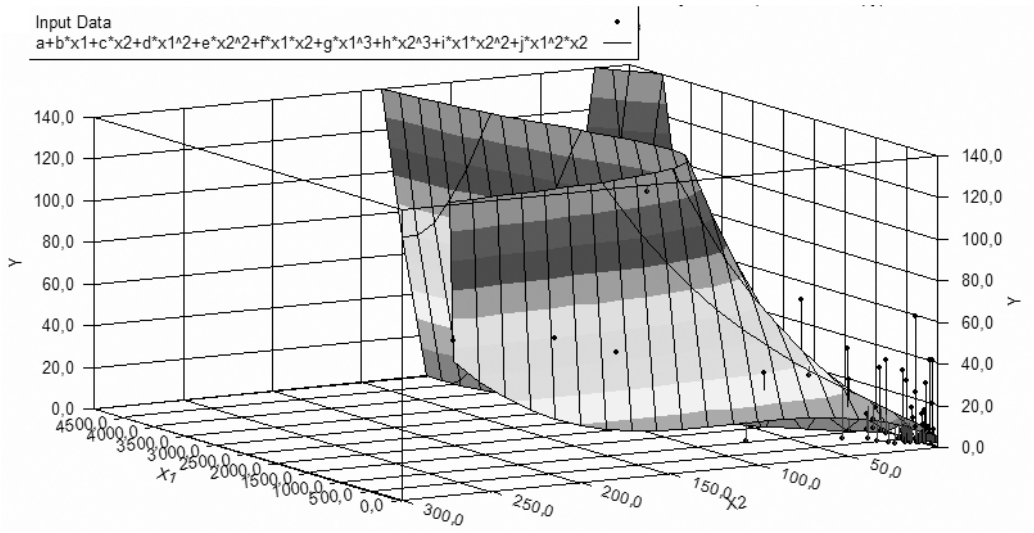

Рисунок 1 – Графическое представление описания данных математической моделью

О качестве построенной модели можно судить, исходя из следующих характеристик: стандартная ошибка оценки равна 11,2986020522291, коэффициент множественной детерминации равен 0,6452864929, статистика Дублина-Ватсона составляет 1,24157318003058. Остатки при построении данной модели представлены на рис. 3.2.

Оценки и статистики качества данной модели не являются остаточными показателями эффективности применения полученной модели, так как модель является лишь одним из этапов выбора k. Применение подхода исследовалось на 285 выборках. Применение данной модели улучшили время выполнения этапа построения графа в 62,45% случаев. В 37,55% случаев время выполнения ухудшилось.

Время выполнения 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 большем минимально необходимого для соблюдения условия связности, даже если время построения графа уменьшилось.

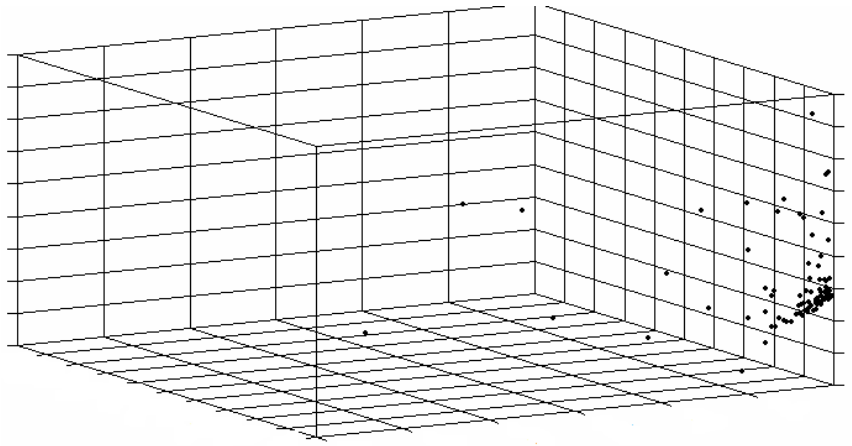

Рисунок 2 – Графическое представление остатков

Сравнение времени выполнения до и после применения модели в зависимости от количества элементов в обрабатываемой выборки представлено на рис 3, и в зависимости от полученного значения k представлено на рис. 4.

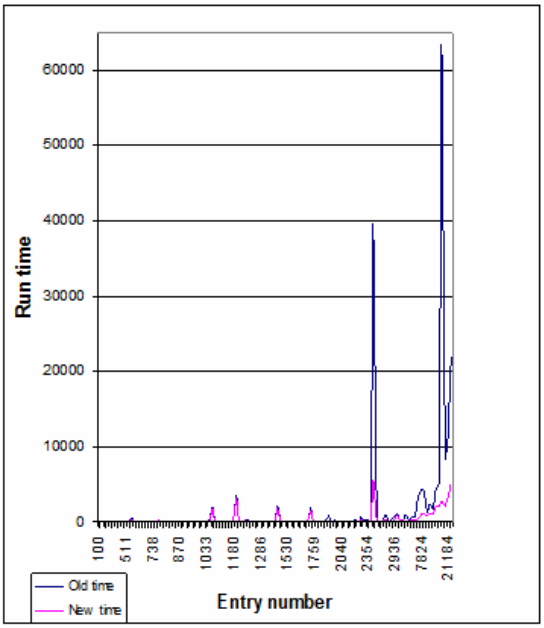

Рисунок 3 – Зависимость времени построения ассиметричного графа в зависимости от количества элементов выборки для модифицированного и не модифицированного вариантов алгоритма

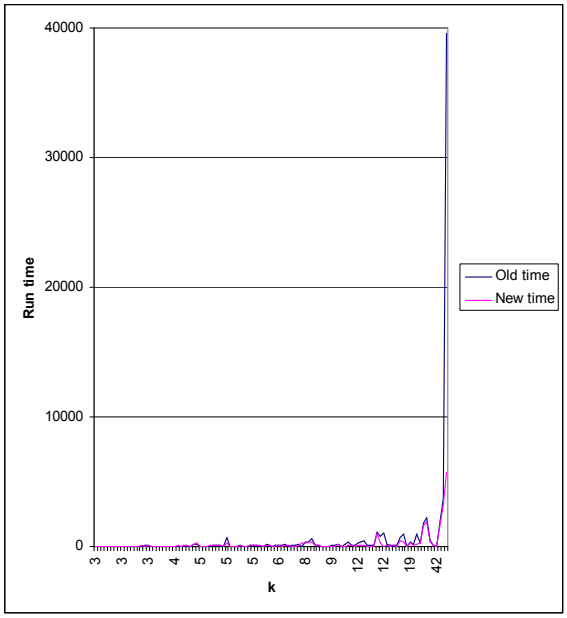

Рисунок 4 – Зависимость времени построения ассиметричного графа в зависимости от полученного значения k для модифицированного и не модифицированного вариантов алгоритма

Так же в результате исследования была построена математическая модель для оптимизации выбора начального значения k при построении симметричного k-nn графа. Модель для ассиметричного k-nn графа имеет следующий вид и представлена на (рис. 5):

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

где x_1- коэффициент расстояния; x_2 – количество компонентов связности. Значения коэффициентов представлены в табл. 2.

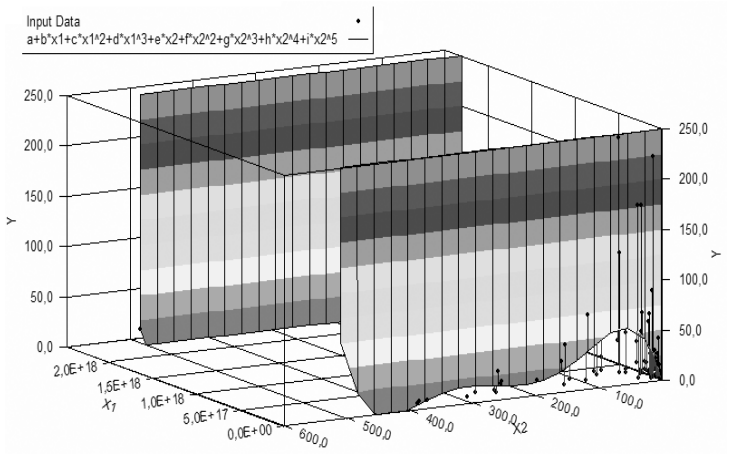

Рисунок 5 – Графическое представление описания данных математической моделью

Таблица 2 – Значения коэффициентов модели для определения k в k-nn графе.

α	-0,547360564	f	-3,09E-02
b	-7,46E-14	g	1,55E-04
c	1,51E-29	h	-3,34E-07
d	-6,56E-48	i	2,61E-10
e	2,323285358		

О качестве построенной модели можно судить, исходя из следующих характеристик: стандартная ошибка оценки равна 42,8805641130193, коэффициент множественной детерминации равен 0,15118817, статистика Дублина-Ватсона составляет 1,26055939255469. Остатки при построении данной модели представлены на рис. 6.

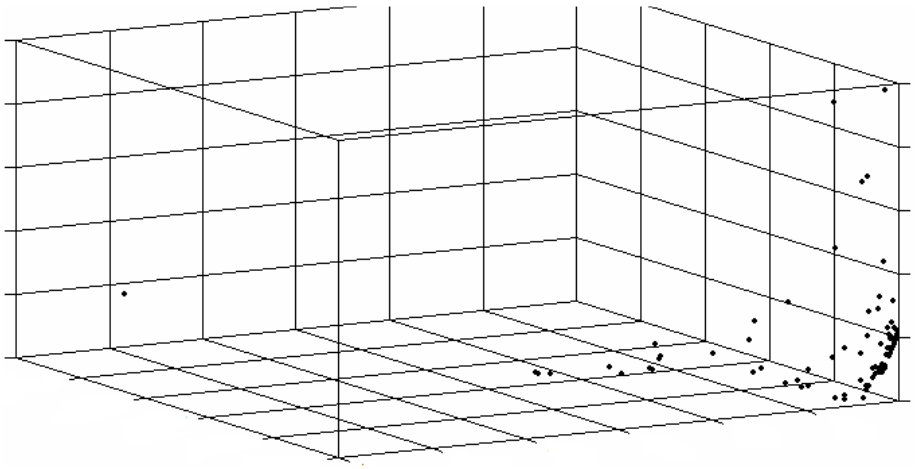

Рисунок 3.6 – Графическое представление остатков

Применение данной 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%.

Сравнение времени выполнения до и после применения модели в зависимости от количества элементов в обрабатываемой выборки представлено на рис 7, и в зависимости от полученного значения k представлено на рис. 8.

Использование модели особенно критично для больших выборок. Полученные результаты будут использованы для дальнейших исследований и модификаций алгоритма Хамелеон.

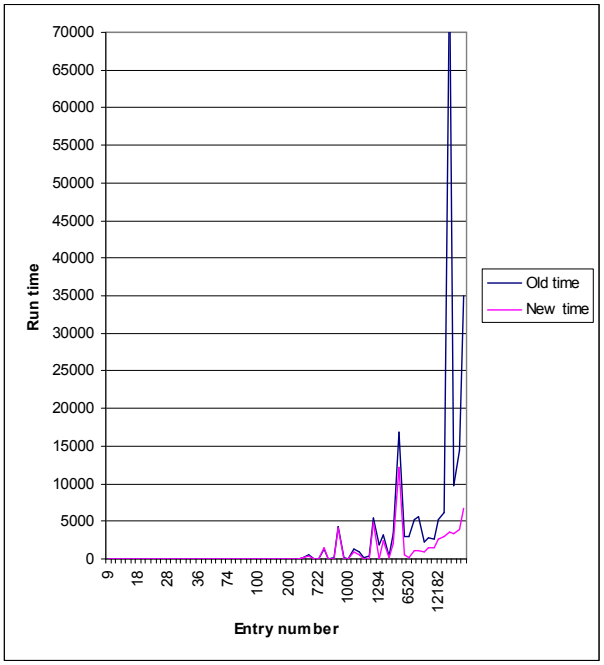

Рисунок 7 – Зависимость времени построения симметричного графа в зависимости от количества элементов выборки для модифицированного и не модифицированного вариантов алгоритма

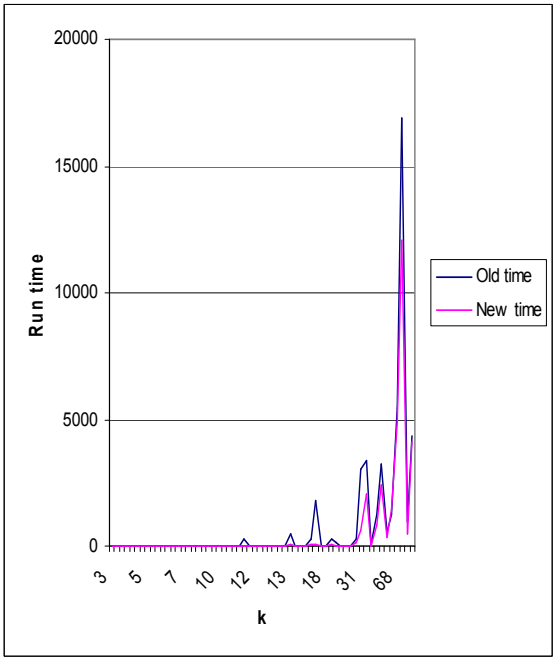

Рисунок 8 – Зависимость времени построения симметричного графа в зависимости от полученного значения k для модифицированного и не модифицированного вариантов алгоритма

4. Математическая модель выбора алгоритмов

Управляемые параметры – характеристики выборки (максимальные и минимальные значения математического ожидания, дисперсии, разброса и вычисляемый параметр).

Составляющими целевого параметра являются:

- алгоритм построения графа;
- мера расстояния;
- алгоритм огрубления графа;
- алгоритм начального разделения графа;
- мера сходства классов;
- алгоритм восстановления графа.

На основании имеющихся алгоритмов составлено 14784 комбинаций для анализа.

Математическая модель, полученная на основании этих данных имеет вид:

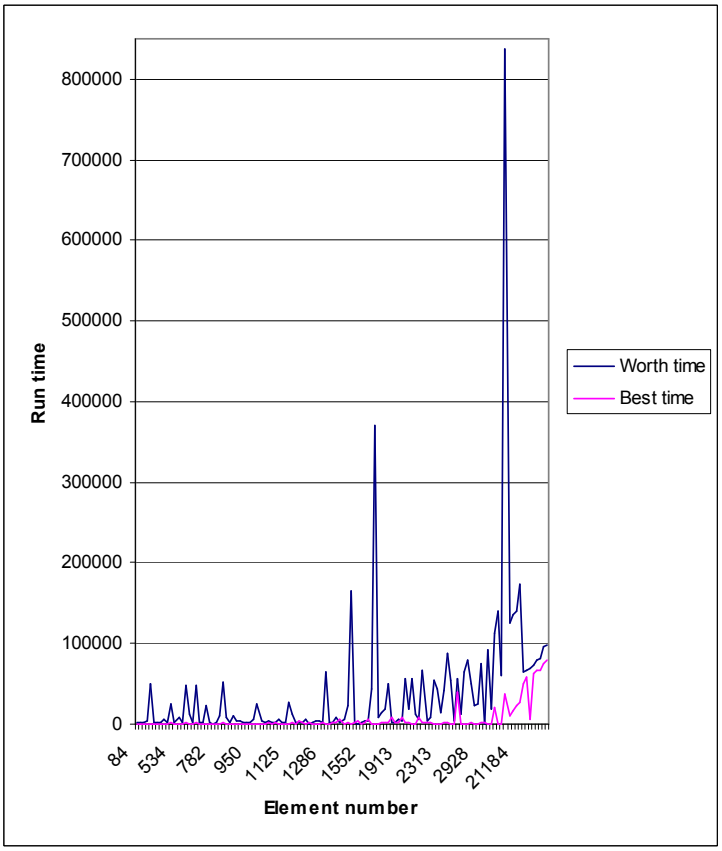

Рисунок 9 – Зависимость времени построения симметричного графа в зависимости от количества элементов выборки для лучшего и худшего вариантов алгоритма.

$$Y = a*x1+b*x2+c*x3+d*x4+e*x5+f*x6+g*x7+h*x8+i*x9+j,$$

где x_1- x_9 соответствуют характеристикам выборок, на основании которых построена модель.

Следует отметить, что в ходе исследования было установлена нецелесообразность использования симметричного алгоритма построения графа и в дальнейших экспериментах будет использован ассиметричный алгоритм. Построение математической модели проводилось на основании результатов экспериментов разделения экспериментальных выборок с помощью различных комбинаций алгоритмов в рамках модифицированного алгоритма Хамелеон.

Выводы

В данной статье разработана математическая модель для выбора алгоритмов в рамках модифицированного алгоритма Хамелеон, построена математическая модель для выбора k при построении k-nn графа в рамках модифицированного алгоритма Хамелеон, приведены результаты применения разработанных методов на реальных данных.

Была разработана математическая модель для выбора k при построении k-nn графа в рамках модифицированного алгоритма Хамелеон. Приведены результаты, полученные как на экспериментальных, так и на реальных данных.

Результаты работы позволили усовершенствовать этап построения графа путем модификации алгоритма Хамелеон с целью улучшения процессов кластеризации, ориентированных на работу с очень большими базами данных.

Список использованной литературы

1. Чубукова И.А. Data Mining БИНОМ / А.И. Чубукова // Лаборатория знаний. Интернет-университет информационных технологий. – ИНТУИТ.ру. – 2008.
2. Brian Read Advances in Databases : 18th British National Conference on Databases, BNCOD 18 Chilton, UK, July 9 – 11. – 2001.
3. Karypis G. Multilevel k-way Partitioning Scheme for Irregular Graphs / G. Karypis, V. Kumar // Journal of parallel and distributed computing. – 1998. – № 48. – P. 96-129
4. Karypis G. A fast and highly quality multilevel scheme for partitioning irregular graphs / G. Karypis, V. Kumar // SIAM J. Sci. Comput., to appear. [Also available on WWW at URL http://www.cs.umn.edu/_karypis]. – 1995.
5. Karypis G. Multilevel k-way Partitioning Scheme for Irregular Graphs / G. Karypis, V. Kumar // Society of Industrial and Applied mathematics. – 1999.
6. Karypis G. Chameleon: Hierarchical Clustering Using Dynamic Modeling / G. Karypis, E.-H. (Sam) Han, V. Kumar // Computer. – 1999. – Vol. 32, № 8. – P. 68-75.

Информация об авторах

Шатовская Татьяна Борисовна – к.т.н. доцент кафедры Программной инженерии, Харьковский Национальный университет радиоэлектроники.
Каменева Ирина Витальевна – к.т.н. старший преподаватель кафедры Программной инженерии, Харьковский Национальный университет радиоэлектроники.