УДК 007.52

А.В. Ляховец
ХНУРЭ, Харьков, Украина, alena-vl@yandex.ru

МОДИФИКАЦИЯ АЛГОРИТМА ВЫБОРА *K* ПРИ ПОСТРОЕНИИ ГРАФА, ИСПОЛЬЗУЮЩЕГОСЯ В АЛГОРИТМЕ КЛАСТЕРИЗАЦИИ ХАМЕЛЕОН

В статье представлена модификация алгоритма выбора *k* для построения асимметричного и симметричного графов в рамках алгоритма Хамелеон. Алгоритм Хамелеон состоит из следующих этапов: построение графа, огрубление, разделение и восстановление. Главной целью данной работы является исследование и улучшение этапа построения графа посредством оптимизации алгоритма выбора *k* при построении графа *k* ближайших соседей. Разработанные математические модели позволят ускорить процесс построения симметричного и асимметричного графов посредством выбора *k* на основании характеристик исходных данных.

КЛАСТЕРИЗАЦИЯ, АЛГОРИТМ ХАМЕЛЕОН, ПОСТРОЕНИЕ ГРАФА, СВЯЗНОСТЬ, *K*-БЛИЖАЙШИХ СОСЕДЕЙ, СИММЕТРИЧНЫЙ ГРАФ, АСИММЕТРИЧНЫЙ ГРАФ

Введение

На данный момент весьма активно исследуются различные методы кластеризации. Каждый метод применим для определенных выборок, и в результате могут быть получены совершенно разные разбиения исходного множества. Выбор определенного метода зависит от типа желаемого результата и характеристик исходного множества. 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* [1-3]. Хамелеон — это новый иерархический алгоритм, который преодолевает ограничения существующих алгоритмов кластеризации. Данный алгоритм рассматривает динамическое моделирование в иерархической кластеризации [4, 5].

В алгоритме можно выделить следующие этапы: построение графа, огрубление, разделение, восстановление и улучшение.

Хамелеон представляет объекты посредством часто используемого графа *k*-ближайших соседей (*k*-nearest neighbor graph, *knn*) [6] В модифицированном алгоритме Хамелеон на этапе построения графа применяются симметричный и асимметричный *knn* графы [7].

1. Определение *k* при построении *k-nn* графа

При решении поставленной задачи при построении графа *k* должно быть выбрано таким образом, чтобы соблюдалось условие связности построенного графа. Граф называется связным, если в нем для любых двух вершин имеется маршрут, соединяющий эти вершины. На практике применяется два принципиально различных порядка обхода, основанных на поиске в глубину и поиске в ширину соответственно.

При поиске в ширину вначале все вершины помечаются как новые, после поочередно для каждой из инцидентных вершин оцениваются все вершины, инцидентные выбранной. Если все инцидентные вершины посещены и остались непосещенные вершины — то граф несвязный.

Главное отличие поиска в глубину от поиска в ширину состоит в том, что при поиске в глубину в качестве вершины для исследования выбирается та из посещенных вершин, которая была посещена последней. Общая оценка трудоемкости для алгоритмов одинаковая — $O(m+n)$.

В каждом из алгоритмов есть возможность оперировать не вершинами, а ребрами, связывающими вершины. Существует два варианта алгоритма поиска в глубину: поиск в глубину с вычислением глубинных номеров — рекурсивный и итеративный варианты. Но так как в данном случае вычисление глубинных номеров не является необходимым, то эти алгоритмы не являются актуальными. Приведенные алгоритмы поиска в глубину и ширину также могут быть реализованы с использованием рекурсии, но в данном случает есть несколько ограничений и недостатков [5].

Общая сложность алгоритма может быть приблизительно оценена значением $\frac{N^3}{8}$. Возможное количество вершин графа ограничено только максимальным размером линейного массива (32 000) [8].

Таким образом, значение *k* последовательно увеличивается, пока граф не станет связным. Так как данная операция трудоемка и длительна, она нуждается в оптимизации.

2. Создание экспериментальных выборок

Для проверки работоспособности метода необходимо большое количество выборок. Для проведения экспериментов было отобрано 47 реальных выборок. Так как при использовании различных входных данных с определенными статистическими характеристиками производительность и качество кластеризации могут сильно отличаться, то необходимо проводить анализ и на синтетических выборках, созданных специально для данной задачи. Исследований в данной области немного и все крайне специфичны для рассматриваемых задач [9].

В данной работе создание 3D фигур выполняется посредством 3D s max studio. Данное 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. Сгенерировано 28 различных выборок.

Также для экспериментов было выбрано 42 широко используемые сгенерированные выборки, применяемые для кластеризации данных.

3. Анализ различных характеристик выборок

Для оптимизации выбора начального параметра *k* при построении *k-nn* графа необходимо построить математическую модель зависимости *k* от характеристик обрабатываемой выборки. Математическая модель будет построена на основе исследования вышеописанных выборок.

Математической моделью называется формальная система, представляющая собой конечное собрание символов и совершенно точных правил оперирования с этими символами в совокупности с интерпретацией свойств определенного объекта некоторыми символами, отношениями и константами

Будем считать, что зависимости между параметрами задаются в виде следующего набора функций (1):

$$W_i = F(X_1, X_2, ..., X_n, a_1, a_2, ..., a_k), i = (1, m), \quad (1)$$

где W — обозначения целевых параметров; X — обозначения управляемых параметров; a — обозначения неуправляемых параметров; m — число целевых параметров; n — число управляемых параметров; k — число неуправляемых параметров.

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 матожидания, дисперсии и разброса. Зависимости между данными параметрами и значением *k* не выявлено [3].
- Во втором эксперименте в качестве управляемого параметра были выбраны длина наибольшего остовного 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. Вторая характеристика вычисляется следующим образом (2):

$$SetDist = \max\left(\frac{dist(avComponent_i, avComponent_j)}{\max\left(\frac{\max ComponentOstovEdge_{ij}}{ComponentVertexNum_{ij}}\right)}\right), \quad (2)$$

где *avComponent* — центроид компонента связности; *ComponentOstovEdge* — ребро, соединяющее вершины, принадлежащие одному компоненту; *ComponentVertexNum* — количество вершин в компоненте.

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

4. Построение математической модели для оптимизации выбора начального значения *k* при построении асимметричного *k-nn* графа

В результате исследования была получена следующая мат модель(3)(рис. 1):

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

где x_1 — коэффициент расстояния; x_2 — количество компонентов связности. Значения коэффициентов представлены в табл. 1.

Таблица 1

Значения коэффициентов модели для определения *k* в *aknn* графе

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

О качестве построенной модели можно судить, исходя из следующих характеристик. Стандартная ошибка оценки равна 11,2986020522291. Коэффициент множественной детерминации

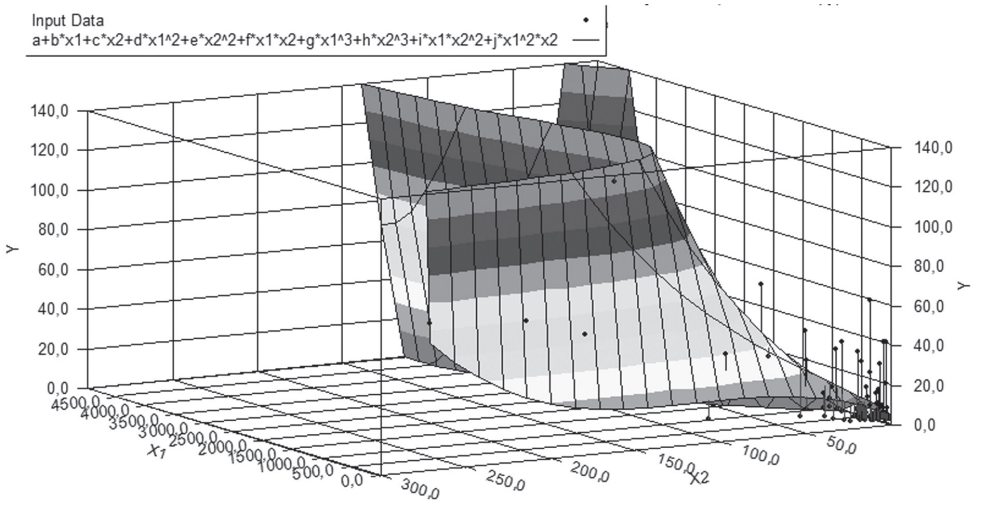

Рис. 1. Графическое представление описания данных математической моделью

равен 0,6452864929. Статистика Дублина-Ватсона составляет 1,24157318003058. Остатки при построении данной модели представлены на рис. 2.

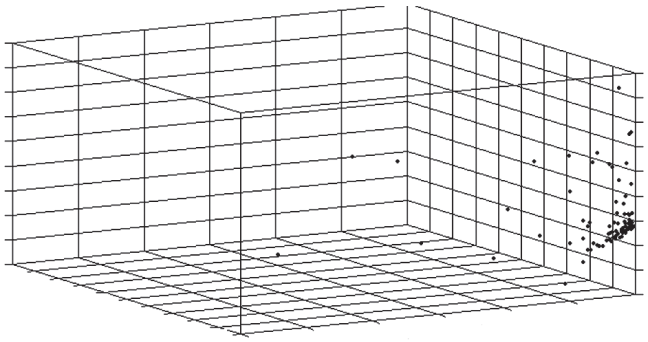

Рис. 2. Графическое представление остатков

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.

5. Построение математической модели для оптимизации выбора начального значения *k* при построении симметричного *k-nn* графа

В результате исследования была получена следующая математическая модель:

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

где x_1 _ коэффициент расстояния; x_2 – количество компонентов связности. Значения коэффициентов представлены в табл. 2. Графическое представление показано на рис. 3.

Таблица 2

Значения коэффициентов модели для определения *k* в *sknn* графе

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

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. 4.

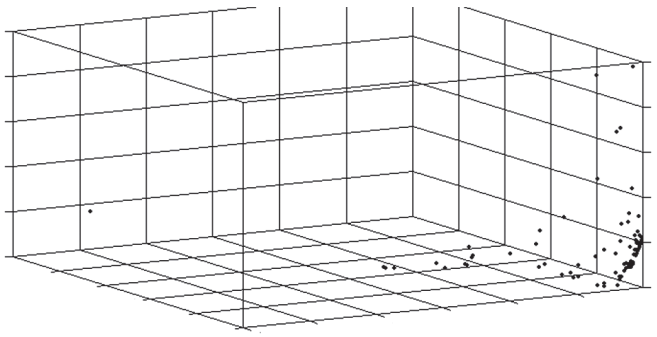

Рис. 4. Графическое представление остатков

Применение данной модели улучшило время выполнения этапа построения графа в 69,23% случаев. В 20,51% случаев время выполнения

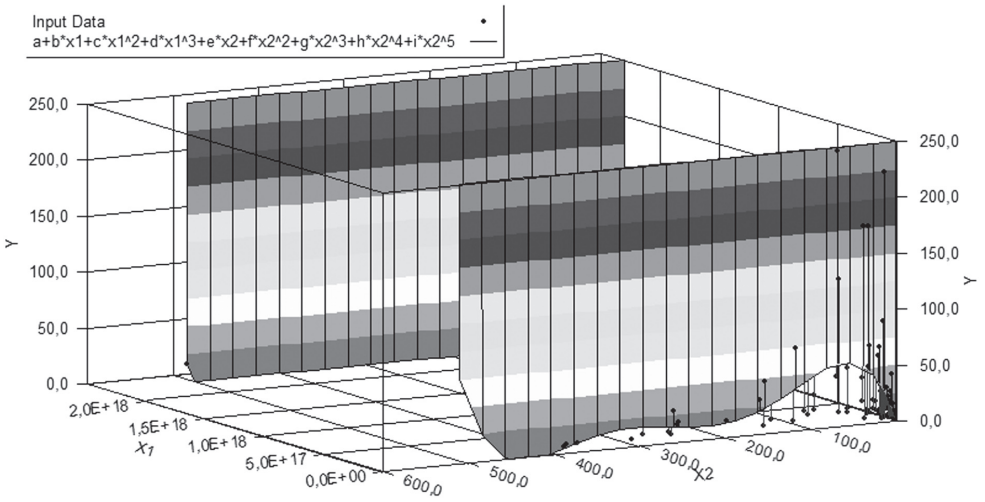

Рис. 3. Графическое представление описания данных математической моделью

ухудшилось. Отрицательный результат применения модели получен в 5,12% случаев. В среднем время выполнения улучшилось на 169%.

Выводы

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 на одном ядре. Время производительности для разных графов может быть использовано для изучения масштабируемости алгоритма.

Использование представленных моделей при построении графа в рамках модифицированного алгоритма Хамелеон позволило сократить время построения графа до 169%. Использование модели особенно критично для больших выборок с большим расстоянием между большими плотными группами объектов. Полученные результаты будут использованы для дальнейших исследований и модификаций алгоритма Хамелеон.

Список литературы: 1. *Чубукова, И.А.* Data Mining [Текст] / И.А. Чубукова // БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий — ИНТУИТ.ру, 2008. **2.** *Matthias Hein and Ulrike von Luxburg* Similarity Graphs in Machine Learning // MLSS 2007 Practical Session on Graph Based Algorithms for Machine Learning August 2007. **3.** *Huan Liu and Nitin Agarwal* Modeling and Data Mining in Blogosphere // (Synthesis Lectures on Data Mining and Knowledge Discovery) (Paperback — Jul 30, 2009). **4.** *George Karypis, Eui-Hong (Sam) Han, Vipin Kumar,* "Chameleon: Hierarchical Clustering Using Dynamic Modeling" // Computer, vol. 32, no. 8, pp. 68-75, Aug. 1999, doi:10.1109/2.781637. **5.** *Ляховец, А.В.* Исследование эффективности динамической кластеризации линейноразделимых зашумленных данных [Текст] / А.В. Ляховец, Н.С. Лесная, Т.Б. Шатовская // Научно технический журнал «Системы обработки информации»5(86)' 2010 с86-91. **6.** *Lyakhovets Alyona,* «Comparison, research and analysis of predictions lumbar spinal stenosis tendencies built by intellectual methods».// Proceedings of the 5-th Internatioal Conference ACSN-2011 Lviv. **7.** *Ляховец, А.В.* Экспериментальные результаты исследования качества кластеризации разнообразных наборов данных с помощью модифицированного алгоритма Хамелеон [Текст]/ А.В. Ляховец // «Вестник запорожского национального университета» №2, Запорожье, 2011, с. 86-73. **8.** *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, May 2011. **9.** *Коршунов, Ю.М.* Получение многомерной статистической выборки с заданными корреляционными свойствами [Текст] / Ю.М. Коршунов // ISSN 1995-4565. Вестник РГРТУ. Рязань, 2008 Вып. 23.

Поступила в редколлегию 03.08.2012

УДК 007.52

Модифікація алгоритму вибору *k* для побудови графа, використаного у алгоритмі кластеризаіі Хамелеон / А.В.Ляховець // Біоніка інтелекту: наук.-техн. журнал. – 2012. – № 2 (79). – С. 76–79.

В статті представлена модифікація алгоритму вибору *k* для побудови асиметричного та симетричного графа у рамках алгоритму Хамелеон. Алгоритм Хамелеон складається з наступних этапів: побудова графа, огрубіння, поділ та відновлення. Головною метою цієї роботы є дослідження та этапу побудови графа через оптимізацію алгоритму выбору *k* при побудові графа *k* до найближчих сусідів. Розробдені математичні моделі дозволяють прискорити процес побудови асиметричного та симетричного графів через выбір *k* на основі характеристик начальних даних.

Табл. 2. Іл. 4. Бібліогр.: 9 найм.

UDC 007.52

***K* selection algorithm modification used for graph building in Xameleon clustering algorithm** / A.V.Lyakhovets //Bionics of Intelligense: Sci. Mag. – 2012. – № 2 (79). – P. 76–79.

I the article *k* selection algorithm is presented. Algorithm is for building symmetrical and asymmetrical graph in frames of Xameleon algorithm. Algorithm Xameleon containes of next steps: graph build, coursing, partitioning and refinement. The main goal of the work is research and modification of graph build step. This step is modified with *k* selection algorithm optimization while *k* nearest neighbors graph build. Developed math models allows speed up the process of symmetrical and asymmetrical graph build by selection of k on the base of initial data set characteristics.

Tab. 2. Fig. 4. Ref.: 9 items.