УДК 519.7:007:004

А.В. ЛЯХОВЕЦ, м.н.с., ХНУРЭ, Харьков

ИССЛЕДОВАНИЕ ЗАВИСИМОСТИ ЗНАЧЕНИЯ *k* ПРИ ПОСТРОЕНИИ *k-nn* ГРАФА ОТ РАЗЛИЧНЫХ ХАРАКТЕРИСТИК ВЫБОРКИ ДЛЯ МОДИФИКАЦИИ АЛГОРИТМА ХАМЕЛЕОН

В статье представлены результаты анализа и экспериментов применения различных характеристик множества для выявления зависимости значения *k* при построении *k-nn* графа от выделенных характеристик выборки. Данная зависимость будет применена в модифицированном алгоритме Хамелеон для ускорения работы алгоритма на этапе построения графа и улучшения качества кластеризации посредством этого ускорения.

Ключевые слова: характеристики выборки, *k-nn* граф, алгоритм Хамелеон, построение графа, кластеризация. Библиогр.: 8 назв.

Постановка проблемы и анализ литературы. На данный момент весьма активно исследуются различные методы кластеризации. В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие базы данных. В них основное внимание уделяется масштабируемости, так как во многих областях за последнее десятилетие существенно выросли объемы данных.

Из анализа работ [1 – 3] можно сделать вывод, что к наиболее актуальным алгоритмам относятся: *BIRCH*, *CURE*, *CHAMELEON*, *ROCK*. В работах [4, 5] более детально описан алгоритм Хамелеон и его применение для кластеризации больших объемов данных [6]. Но на данный момент все еще не решена проблема быстродействия при кластеризации больших объемов данных. Быстродействие алгоритма Хамелеон в целом может быть улучшено посредством повышение скорости работы его на отдельных этапах.

Цель работы – повышение скорости построения графа в алгоритме Хамелеон посредством определения значения *k* на основании характеристик анализируемых данных.

Описание модифицированного алгоритма Хамелеон. Хамелеон – это новый иерархический алгоритм, который преодолевает ограничения существующих алгоритмов кластеризации. Данный алгоритм рассматривает динамическое моделирование в иерархической кластеризации [7].

В алгоритме можно выделить следующие этапы: построение графа, огрубление, разделение, восстановление и улучшение [8].

Хамелеон представляет объекты посредством часто используемого графа *k*-ближайших соседей (*k*-nearest neighbor graph).

Создание экспериментальных выборок. Для проверки влияния той или иной характеристики выборки на значение *k* необходимо большое количество выборок. Отсутствие реального источника данных требуемого объема, разнообразия и качества вынуждает обратиться к альтернативному источнику [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. Добавление шума в выборку производится непосредственно перед проведением анализа. Выборки анализировались в 4 состояниях: без добавления шума, с добавлением 20 %, 40 % и 60 % шума.

Были использованы 54 выборки при проведении эксперимента.

Определение *k* при построении *k-nn* графа. При решении поставленной задачи построения 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.

Рассмотрим итерактивные алгоритмы и алгоритмы, реализованные с помощью рекурсии [10].

В худшем случае (при полном графе) рекурсивный алгоритм, перебирая все возможные ребра, будет вынужден вызвать основную процедуру $(N-1)!$ раз. Велика вероятность, что при достаточно большом N произойдет переполнение оперативной памяти, которое вызовет ошибку. Кроме того, размеры квадратной матрицы смежности дают сильное ограничение на возможное количество вершин графа: не более 250.

Итеративный же алгоритм переберет все ребра графа, которых может быть не более чем $\frac{N\times(N+1)}{2}$. Следовательно, общая сложность алгоритма может быть приблизительно оценена значением $N^3/8$. Возможное количество вершин графа ограничено только максимальным размером линейного массива (32 000).

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

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 30 выборок.

Математической моделью называется совокупность математических соотношений, уравнений, неравенств, описывающих основные закономерности, присущие изучаемому процессу, объекту или системе.

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

$$W_i = F(X_1, X_2, \ldots, X_n, a_1, a_2, \ldots, a_k), \; i = \overline{(1, m)}, \qquad (1)$$

где W_i – обозначения целевых параметров; X_q $\left(q = \overline{1, n}\right)$ – обозначения управляемых параметров; a_p $\left(p = \overline{1, r}\right)$ – обозначения неуправляемых параметров; m – число целевых параметров; n – число управляемых параметров, значения которых можно выбирать в технически допустимых пределах и тем самым влиять на процесс моделирования; r – число неуправляемых параметров.

Так как построенная математическая модель должна отображать зависимость между начальным значением параметра k при построении k-nn графа и характеристиками выборки, то в соответствии с формулой (1) целевым параметром является значение k, а вычисляемые характеристики выборки являются управляемыми параметрами.

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

1. 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].
2. Во втором эксперименте в качестве управляемого параметра были выбраны длинна наибольшего остовного ребра полносвязного графа и среднее значение длинны всех остальных ребер остова. Данные характеристики показывают зависимость, но использование этого

подхода не является целесообразным в связи с трудоемкостью построения остова полносвязного графа.

3. В третьем эксперименте в качестве характеристики использовались количество объектов в выборке и максимальное расстояние между компонентами связности за вычетом средних значений. Данные характеристики нетрудоемки в расчете и существует зависимость между ними и значением k. Наличие такой зависимости позволит построить математическую модель рассчета параметра k для ускорения построения графа в модифицированном алгоритме Хамелеон. Уменьшение времени работы на этапе построения графа сократит общее время работы алгоритма.

Выводы. Характеристики, основанные на компонентах связности, могут быть использованы для построения математической модели зависимости k от характеристик выборки. Полученные результаты будут использованы для дальнейших исследований и модификаций алгоритма Хамелеон.

Список литературы: **1.** *Чубукова И.А.* Data Mining / *И.А. Чубукова.* – БИНОМ. Лаборатория знаний, Интернет-университет информационных технологий – ИНТУИТ.ру, 2008. – С. 384. **2.** *Hein M.* Similarity Graphs in Machine Learning MLSS / *M. Hein, U. Luxburg* // Practical Session on Graph Based Algorithms for Machine Learning August 2007. – P. 22. **3.** *Liu H.* Modeling and Data Mining in Blogosphere / *H. Liu, N. Agarwa* // Synthesis Lectures on Data Mining and Knowledge Discovery Paperback. Jul 30, 2009. – P. 109. **4.** *Karypis G.* Chameleon: Hierarchical Clustering Using Dynamic Modeling / *G. Karypis, E.S. Han, V. Kumar* // Computer. – 1999. – Vol. 32. – №. 8 – P. 68-75. **5.** *Бувайло Д.П.* Быстрый высокопроизводительный алгоритм для разделения нерегулярных графов / *Д.П. Бувайло, В.А. Толок* // Вісник Запорізького державного університету – 2002 – № 2. – С. 47-53. **6.** *Кузнецов Д.Ю.* Кластерный анализ и его применение / *Д.Ю. Кузнецов, Т.Л. Трошина* // Ярославский педагогический вестник. – 2006. – №. 4. – С. 103 107 **7.** *Ляховец А.В.* Исследование эфективности динамической кластеризации линейноразделимых зашумленных данных / *А.В. Ляховец, Н.С. Лесная, Т.Б. Шатовская* // Научно технический журнал "Системы обработки информации". – 2010 – 5 (86) – С. 86-91. **8.** *Agarwal P.* Challenges and Tools of Clustering Algorithms / *P. Agarwal, A M. Afshar, R. Biswas* // IJCSI International Journal of Computer Science Issues – 2011 – Vol. 8, Issue 3. – №. 2. – P. 79-81. **9.** *Коршунов Ю.М.* Получение многомерной статистической выборки с заданными корреляционными свойствами / *Ю.М. Коршунов* // Вестник РГРТУ. – 2008 – №.23. **10.** SPARCL: Efficient and Effective Shape-Based Clustering / *V. Chaoji, M.A. Hasan, S. Salem, M.J. Zaki* // In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy. IEEE Computer Society – P. 93-102.

Статью представил д.т.н., проф. ХНУРЭ Четвериков Г.Г.

УДК 519.7:007:004

Дослідження залежності значення k при побудові k-nn графу від різних характеристик вибірки для модифікації алгоритма Хамелеон. / Ляховець А.В. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2012. – № 62 (968). – С. 130 – 134.

У роботі представлено результати аналізу та експериментів що до використання різних характеристик множини для визначення залежності значення *k* при побуджові *k-nn* графу від обраних характеристик. Ця залежність буде використана у модифікованому алгоритмі Хамелеон для прискорення роботи алгоритму на этапі побудови графу та через це, прискорення та поліпшення якості кластеризації.

Ключові слова: характеристики вибірки, *k-nn* граф, алгоритм Хамелеон, побудова графу, кластеризація. Бібліогр.: 10 назв.

UDC 519.7:007:004

Research of dependency k while k-nn graph building from different data set characteristics for Khameleon algorithm modification. / Lyakhovets A.V. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2012. – №. 62 (968). – P. 130 – 134.

In the article research and analysis results are presented. The main point was to use different data set characteristics for finding dependence between *k* value which is used for *k-nn* graph build and this characteristics. This dependence will be used in modification of Chameleon algorithm for graph build stage acceleration and by this, clustering acceleration and quality improvement. Refs.: 10 titles.

Keywords: dataset characteristics, *k-nn* graph, Chameleon algorithm, graph build, clustering.

Поступила в редакцию 18.07.2012