ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 004.9

САМООРГАНИЗУЮЩАЯСЯ КЛАСТЕРИЗАЦИЯ ПОТОКА БОЛЬШИХ ДАННЫХ

Статья поступила в редакцию 06.02.2020, в окончательном варианте – 26.02.2020.

Печеный Евгений Абрамович, Казанский национальный исследовательский технологический университет, 420015, Российская Федерация, Республика Татарстан, г. Казань, ул. К. Маркса, 68,

кандидат технических наук, доцент, РИНЦ AuthorID 408103, e-mail: platova51@mail.ru

Нуриев Наиль Кашапович, Казанский национальный исследовательский технологический университет, 420015, Российская Федерация, Республика Татарстан, г. Казань, ул. К. Маркса, 68,

доктор педагогических наук, кандидат технических наук, профессор, заведующий кафедрой информатики и прикладной математики, ORCID https://orcid.org/0000-0002-9557-5493, РИНЦ AuthorID 527783, e-mail: nurievnk@mail.ru

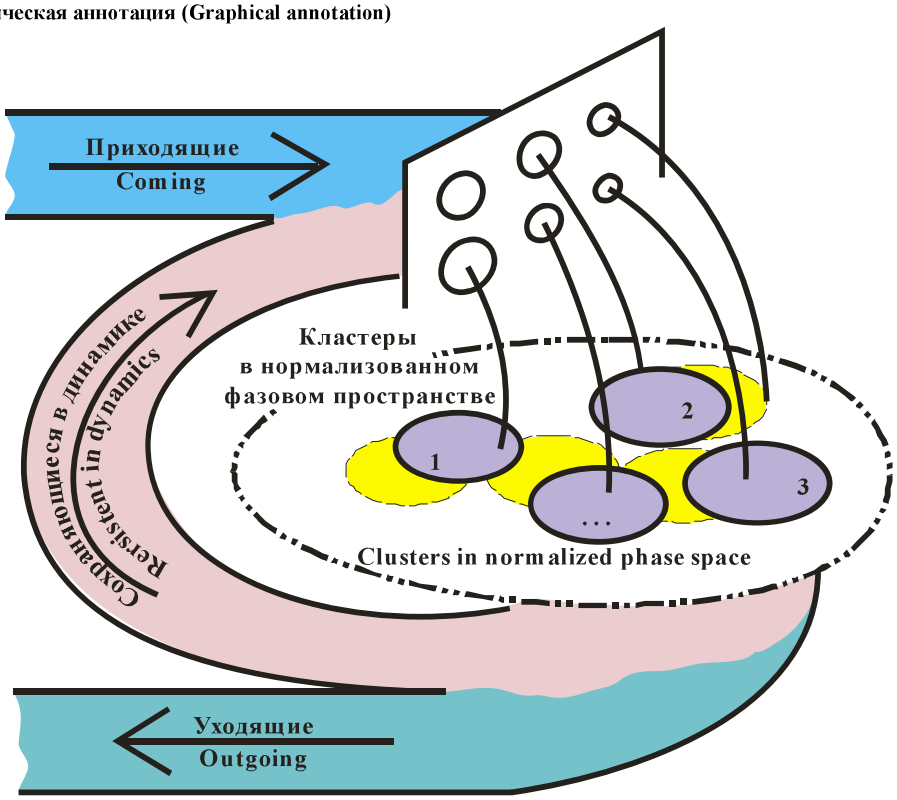
Старыгина Светлана Дмитриевна, Казанский национальный исследовательский технологический университет, 420015, Российская Федерация, Республика Татарстан, г. Казань, ул. К. Маркса, 68,

кандидат педагогических наук, доцент, ORCID https://orcid.org/0000-0002-3401-6452, РИНЦ AuthorID 237000, e-mail: svetacd_kazan@mail.ru

В работе представлена математическая модель и дано описание алгоритма на основе аппарата кластерного анализа, ориентированного на проведение процедур классификации «больших данных». В качестве кластеров предложено использовать сфероиды, для построения которых предварительно производится нормирование переменных и преобразование их в безразмерную форму. Простота аналитического описания формы кластеров служит эффективной защитой алгоритма от «проклятия размерности», обеспечивает сохранение его работоспособности при большом числе классифицируемых признаков. Отличительной особенностью разработанного алгоритма является его способность функционировать в динамическом режиме, т.е. в условиях изменений свойств объектов, присутствующих в кластерах; пополнения кластеров потоком новых объектов; удаления некоторых объектов из числа классифицируемых. Для обеспечения однозначности выделяемых классификационных категорий в алгоритме предусмотрена защита от пересечений кластеров. Важным и полезным эксплуатационным качеством алгоритма является его самоорганизуемость. Он может обрабатывать данные в потоке без участия оператора, выполняя, по мере необходимости, коррекцию положения и размеров кластеров. Процедура коррекции представляет собой последовательность итераций, в ходе которой осуществляется сближение геометрических центров кластеров с центрами группировок объектов, имеющихся в их составе. В статье приведена блок-схема алгоритма, который был реализован программно. Работа этого алгоритма продемонстрирована и графически проиллюстрирована на примере сравнительно небольшого массива данных, элементы которого описываются двумя классифицирующими признаками.

Ключевые слова: большие данные, классификация объектов, кластеры, динамическая кластеризация, самоорганизация, самообучение

Графическая аннотация (Graphical annotation)



SELF-ORGANIZING CLUSTERING OF GREAT DATA FLOW

The article was received by the editorial board on 06.02.2020, in the final version – 26.02.2020.

Pechenyy Eugene A., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420015, Republic of Tatarstan, Russian Federation,

Cand. Sci. (Engineering), Associate Professor, RISC AuthorID 408103, e-mail: platova51@mail.ru

Nuriev Nail K., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420015, Republic of Tatarstan, Russian Federation,

Doct. Sci. (Pedagogy), Cand. Sci. (Engineering), Professor, Head of the Department of Informatics and Applied Mathematics, ORCID https://orcid.org/0000-0002-9557-5493, RISC AuthorID 527783, e-mail: nurievnk@mail.ru

Starygina Svetlana D., Kazan National Research Technological University, 68 Karl Marks St., Kazan, 420015, Republic of Tatarstan, Russian Federation,

Cand. Sci. (Pedagogy), Associate Professor, ORCID https://orcid.org/0000-0002-3401-6452, RISC AuthorID 237000, e-mail: svetacd_kazan@mail.ru

This paper presents a mathematical model and provides a description of an algorithm, based on clustering apparatus focused on carrying out big data classification procedures. Spheroids are proposed to be used as clusters, for which to be constructed the variables are preliminarily normalized and transformed to be nondimensional. Simplicity of analytically defining the forms of clusters serves for the algorithm as an efficient protection from the curse of dimensionality and makes it efficient for a great number of criteria to be classified. The distinctive feature of the algorithm developed is its ability to function dynamically, i. e., in the conditions of changing properties of elements available in clusters and refilling the clusters with the stream of new elements. To ensure the unambiguity of the classification categories distinguished, the algorithm provides protection from cluster intersections. An important and useful operating characteristic of the algorithm is its self-organizability. It can process stream data without the intervention of an operator, correcting the locations and sizes of clusters, where necessary. Correcting procedure represents a sequence of iterations, in which the geometric centers of clusters approach to the centers of groups of objects available in the clusters. The paper presents a control-flow chart that was implemented in software. The algorithm operation is demonstrated and illustrated graphically, exemplified by a comparatively small data array, the elements of which are defined by two classification criteria.

Key words: big data, object classification, clusters, dynamic clustering, self-organization, self-study

Стремительное развитие информационных технологий и интернета, которым ознаменовался конец XX – начало XXI в., сделало возможным формирование и передачу очень больших объемов информации. Это, в свою очередь, привело к необходимости создания специальных методов, позволяющих быстро обрабатывать массивы так называемых «больших данных» [1, 2, 6, 9, 16, 17, 20]. Особый интерес и актуальность имеют задачи, связанные с обработкой массивов больших данных в процессе их динамических изменений. Действительно, полностью сложившиеся, неизменяемые большие объемы информации сейчас можно отыскать только в архивных хранилищах и каталогах. Массивы больших данных, являясь, по сути дела, выборками каких-то генеральных совокупностей, находятся в процессе постоянного обмена элементами друг с другом. Иначе говоря, существует поток (потоки), пополняющий массив новыми элементами, и поток уноса (гибели) части элементов массива, что влечет за собой изменения его численного состава. Изменения могут происходить также с признаками, характеризующими состояние элементов. Следовательно, «большие данные» представляют собой сложные динамические образования, подверженные как качественным, так и количественным изменениям. При разработке алгоритмов, предназначенных для работы с большими данными, это обстоятельство должно быть обязательно учтено.

Остановимся более подробно на понятии «обработка данных», которое является весьма широким и нуждается, по мнению авторов, в некоторой конкретизации. Этот термин может быть с достаточным основанием применен к ряду разнородных процедур: устранению недостоверных результатов и исправлению неточностей, пакетированию данных, классификации элементной базы больших данных по определенным признакам и т.п. Здесь мы будем понимать под обработкой только процедуру классификации по совокупности количественно измеримых признаков.

Обратим внимание на довольно существенные различия классификационных техник для качественных и количественных признаков. Например, классификация социума по признакам пола, расовой или национальной принадлежности, группе крови, как правило, дает однозначный результат и не вызывает затруднений. Решение же задачи классификации по фактору возраста далеко не так очевидно и зависит не только от состава данных, но и от целей исследования. Где следует установить границы молодости, зрелости и старости; как будет происходить изменение этих границ в течение 10, 20, 50 лет, в том числе при изменениях среднего возраста населения стран; достаточно ли для вынесения обоснованных суждений этих трех классификационных областей, или целесообразно введение дополнительных – эти вопросы не имеют готовых ответов, пригодных для любых ситуаций. Скорее всего, мнение специалистов-медиков будет отлично от мнения социологов, поскольку физиологическое развитие человеческой личности не совпадает с изменениями ее социального статуса. Перечисленные особенности процедур классификации при наличии количественно измеримых признаков делает подобные задачи достаточно сложными для формализации, а трудоемкость их решения многократно возрастает с ростом размерности пространства классифицирующих признаков. В настоящей работе для решения задачи классификации на массиве больших данных в условиях динамических изменений последних предлагается использовать алгоритм, в основе которого лежит аппарат кластерного анализа.

Кластерный анализ как инструмент группировки объектов внутри исследуемого массива данных известен давно [4, 5, 7, 10–15, 18] и достаточно широко применяется для решения практических задач. Основная его идея состоит в разбиении множества всех элементов массива на конечное число непересекающихся подмножеств (кластеров), элементы которых близки друг к другу в смысле введенной метрики. Это позволяет в ряде случаев отождествить свойства отдельных объектов со свойствами кластера, которому они принадлежат, и заменить решение нескольких частных задач решением одной общей задачи по единому для всех элементов кластера критерию. Такой подход часто и успешно используется при решении проблем логистики, вопросов энергоснабжения и водоснабжения, управления городским хозяйством и пр. Практическое применение кластерного анализа ограничивается отсутствием универсальных рекомендаций по выбору формы кластеров, их размеров и метрики. Эти вопросы решаются, как правило, с использованием эвристических методов, а результаты в значительной мере зависят от интуиции и опыта лица, принимающего решение.

Рассмотрим массив данных, элементы которого характеризуются n количественно измеримыми признаками. Мощность массива Q не является постоянной величиной, а меняется под влиянием двух потоков: пополнения и гибели. Объекты, остающиеся в составе массива, также могут меняться под влиянием изменений одного или нескольких связанных с ними признаков. Таким образом, рассматриваемый массив является динамическим образованием, структура и свойства которого есть явные функции времени.

Если, как упоминалось выше, все признаки, характеризующие состояние объектов Q, количественно измеримы, то геометрическим образом любого объекта будет точка $x^j = (x_1^j,\ x_2^j, ..., x_n^j)$ для $\forall\, j \in N$. Тогда, используя в качестве меры близости эвклидову метрику, получим

$$\left\| x^j - x^k \right\| = \sqrt{\sum_{i=1}^{n} (x_i^j - x_i^k)^2} \quad j,k \in N, \tag{1}$$

где n – размерность пространства классифицирующих признаков. В большинстве случаев разнообразие форм кластеров ограничивается классом выпуклых множеств. Иначе говоря, в рамках традиционных метрик понятие расстояния как меры близости между объектами утрачивает свою определенность. Однако даже при выполнении условия выпуклости с увеличением размерности пространства классифицирующих признаков резко возрастает сложность аналитического описания кластерных границ и интерпретация полученных результатов. Именно по этим причинам авторы остановили свой выбор на кластерах сферической формы, поскольку сложность математического описания сферы как геометрического объекта практически не зависит от размерности пространства, а значит, эта форма как нельзя более подходит для работы с большими данными.

Построение кластерных сфероидов непосредственно в пространстве классифицирующих признаков, конечно, невозможно. Признаки, описывающие состояние объектов массива данных, могут иметь различную природу, разные единицы измерений, а потому понятие «тело» или «фигура» в пространстве таких признаков неопределенно. По этой причине необходимо для всех классифицирующих признаков осуществить переход к новым переменным, измеряемым в единых безразмерных единицах. Поскольку массив данных в течение всего времени существования претерпевает динамические изменения, минимальные и максимальные значения классифицирующих признаков его элементов не могут быть зафиксированы и использованы для перехода к безразмерным величинам. Вследствие этого путем анализа априорной информации и физических особенностей признаков устанавливаются точная верхняя $x_{\sup}^j$ и точная нижняя $x_{\inf}^j$ грани для каждого из n признаков, после чего выполняется переход к безразмерным переменным y^j по формуле

$$y^j = M \frac{x^j - x_{\inf}^j}{x_{\sup}^j - x_{\inf}^j}, \tag{2}$$

где M – масштабирующий множитель, выбираемый из соображений удобства представления данных. Далее в тексте множество векторов $\left\{ y^j \right\}\ j \in N$ будем называть модифицированными значениями классифицирующих признаков.

Рассмотрим последовательно все этапы построения алгоритма кластеризации модифицированного массива данных в режиме динамических изменений. Он позволяет осуществить классификацию массива данных с большим числом классифицирующих признаков в самоорганизующемся режиме.

На множестве модифицированного массива данных $\left\{ y^j \right\}$ выделим некоторое подмножество образов (точек), образующих более или менее тесную группировку, и построим кластерный сфероид $S^1(r, e_0^1)$ с центром в точке e_0^1 и радиусом r. Решения, принимаемые на этом этапе, являются исключительно эвристическими, поскольку формализованные рекомендации по выбору количества кластеров, положениям их центров, величинам радиусов дать невозможно. Поэтому выбор определяется сообразно целям исследования и во многом зависит от квалификации и субъективных предпочтений исследователя. Как бы то ни было, в состав кластера № 1 $S^1(r, e_0^1)$ войдут все элементы модифицированного массива, удовлетворяющие условию $\left\| e_0^1 - y^j \right\| \le r$.

Положение точки e_0^1, которая является геометрическим центром кластера $S^1(r, e_0^1)$, может совпасть с центром группировки (центром масс) элементов, находящихся в границах кластерного сфероида, только случайно. Однако именно близость к центру группировки есть главное классификационное условие. Обозначим начальное положение центра группировки кластера № 1 как g_0^1. Оно легко находится по формуле

$$g_0^1 = \frac{1}{\left| S^1(r, e_0^1) \right|} \sum_{y \in S^1(r, e_0^1)} y^j, \tag{3}$$

где $\left|S^1(r, e_0^1)\right|$ – мощность подмножества элементов в составе кластера $S^1(r, e_0^1)$. Перемещая геометрический центр кластера в точку g_0^1, получим новый сфероид $S^1(r, e_1^1)$. В результате такого движения в сфероиде $S^1(r, e_1^1)$ появятся элементы массива данных, которых не было в $S^1(r, e_0^1)$, а часть элементов, входящих в состав $S^1(r, e_0^1)$, покинут кластер. Очевидно, что и в сфероиде $S^1(r, e_1^1)$, несовпадение геометрического центра e_1^1 и центра группировки g_1^1 также может иметь место. В соответствии с описанной процедурой вновь перемещаем центр сфероида в точку g_1^1 и получаем новый кластерный сфероид $S^1(r, e_2^1)$. Согласно известной теореме функционального анализа [8], последовательность точек $\{e_h^1\}$ сходится, а значит, за конечное число шагов может быть достигнуто выполнение условия

$$\left\| e_h^1 - e_{h-1}^1 \right\| < \varepsilon, \tag{4}$$

где ε – любое наперед заданное положительно число, определяющее приемлемую меру близости между геометрическим центром кластерного сфероида и центром группировки находящихся в его составе элементов.

В процессе реализации данного алгоритма может оказаться, что величина радиуса кластерной сферы r и часть элементов модифицированного массива, которые должны были попасть в состав кластера $S^1(r, e_h^1)$, в него не вошли, и это не соответствует содержательному смыслу задачи. Тогда радиус сфероида увеличивается так, чтобы самый удаленный от точки e_h^1 объект из тех, которые должны войти в кластер, оказался на его границе. Обозначим $\{y^j\}$ – множество элементов модифицированного массива данных, которое необходимо дополнительно ввести в состав расширенного кластерного сфероида. Величина радиуса увеличенного кластера с центром в точке e_h^1 $S^1(R, e_h^1)$ при этом составит

$$R = \max_{\{y^j\}} \left\| e_h^1 - y^j \right\|. \tag{5}$$

Очевидно, что после изменения размеров кластера и введения в его состав новых элементов положение центра группировки опять сместится относительно точки геометрического центра e_h^1. Однако с помощью описанного выше итерационного алгоритма это смещение устраняется.

Построение других кластерных сфероидов происходит по аналогичному сценарию. Начальные размеры кластеров, их необходимое количество m и положение их геометрических центров e_0^p $p = \overline{1, m}$ задается согласно целям проводимого исследования и особенностям обрабатываемого массива данных. После того как формирование системы кластеров по материалам достаточно представительной обучающей выборки будет завершено, можно приступить к эксплуатации системы в динамическом режиме, то есть в условиях действия потоков пополнения и гибели. Разработанный алгоритм, функционируя в режиме динамической кластеризации, проявляет важное и весьма ценное с практической точки зрения качество: самоорганизуемость. В ходе обмена элементами с внешней средой наполнение кластеров изменяется, и алгоритм без участия пользователя в автоматическом режиме непрерывно выполняет коррекцию положения кластерных сфероидов по заданной величине допустимого отклонения геометрических центров кластеров от центров их группировки. Следует заметить, что по мере наполнения кластеров смещение центров группировок заметно уменьшается и может не требовать корректирующего воздействия.

Особое внимание при реализации алгоритма следует обратить на возможность возникновения пересечений двух и более кластеров, которые могут появиться в процессе коррекции их положения и размеров. Этого ни в коем случае нельзя допускать, ибо тогда встает проблема неоднозначной классификации объектов, что дискредитирует саму идею использования кластерного анализа в классификационных процедурах. Для предотвращения подобной ситуации после каждой коррекции положения кластеров в алгоритме предусмотрена проверка условия

$$d(e_k^j, e_h^i) \le r_k^j + r_h^i \qquad \forall j \ne i. \tag{6}$$

Его выполнение означает, что расстояние между геометрическими центрами j-го и i-го кластеров оказалось меньше суммы их радиусов. В этом случае радиус одного из кластеров уменьшается до тех пор, пока пересечение не будет устранено. Подробная блок-схема алгоритма представлена на рисунке 1.

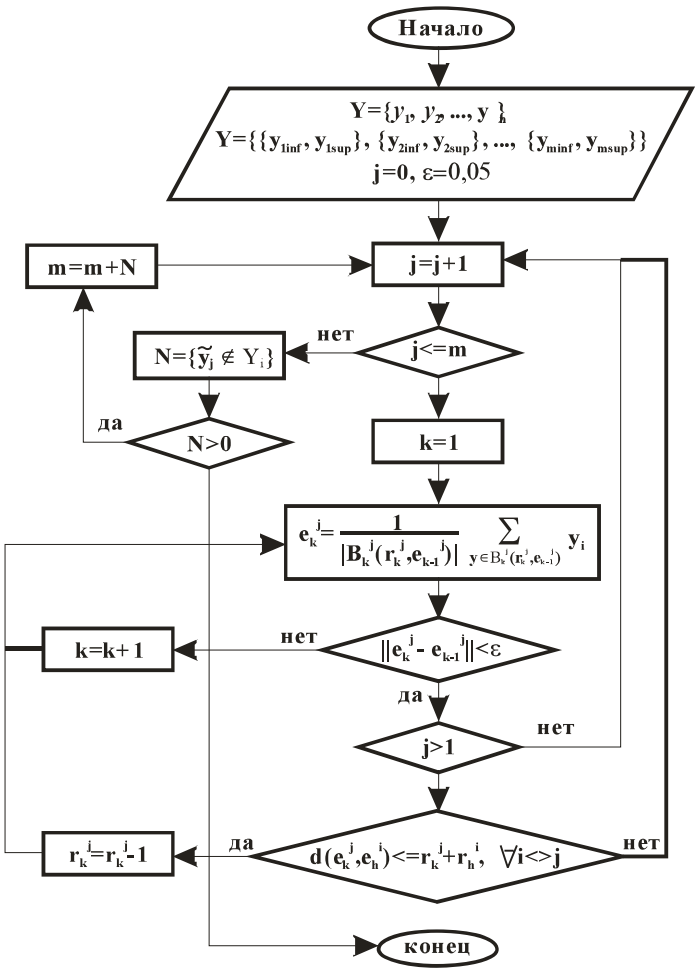


Рисунок 1 – Блок-схема алгоритма динамической кластеризации

Разработанный алгоритм показал высокие эксплуатационные характеристики на массивах больших данных, о чем будет более подробно сказано ниже при рассмотрении примеров его практического применения. Здесь же авторы считают необходимым упомянуть еще об одной особенности своей разработки. Как известно, сферы не могут заполнять пространство полностью. Более того, как указывает М. Гарднер [3], не получен ответ на вопрос о том, какую долю пространства произвольной размерности заполняют сферические тела при наиболее плотном способе упаковки. Отсюда следует, что в ходе построения классификационных категорий какая-то часть массива данных не будет классифицирована, поскольку не войдет в границы ни одного из сформированных кластерных сфероидов. В динамике доля объектов, не прошедших классификацию, может изменяться как в ту, так и в другую сторону, вместе с изменениями положения кластеров, и ликвидировать ее полностью невозможно. Этот недостаток, имманентно присущий рассмотренному алгоритму, обусловлен неизменяемыми свойствами кластеров как геометрических тел. Отмеченный факт, конечно, следует иметь в виду при проведении классификации. Однако он не может оказать сколь-нибудь существенного влияния на результаты классификации и поставить под сомнение их достоверность. С практической точки зрения, действительно, если из миллиона данных массива две или три тысячи объектов не будут классифицированы, то общую картину это никак не изменит.

Для демонстрации возможностей алгоритма и их геометрической иллюстрации был выбран сравнительно небольшой массив (несколько сотен элементов), составленный из городов России, классифицирующими признаками которых является площадь занимаемой ими территории W и численность населения X – эти сведения имеются в интернете и открыты для доступа (например, https://wiki2.org/ru/Список_городов_России_с_территорией_больше_100_квадратных_километров). В качестве точной нижней грани по признаку X, т.е. X_{inf}, была выбрана величина 10000, поскольку населенные пункты с меньшей численностью жителей не имеют в РФ статуса города. В качестве же X_{sup} было взято число 5000000, поскольку городов с населением, превышающим 5 миллионов жителей, в России только 2, и они имеют особый статус и структуру. По тем же основаниям в качестве

точной нижней границы W_{inf} принята площадь 1 км^2, а в качестве точной верхней границы W_{sup} – 1400 км^2, так как город Санкт-Петербург, население которого 5,2 миллиона человек, занимает именно такую территорию.

Применяя формулу (2), преобразуем признаки X и W к безразмерным величинам с масштабирующим множителем $M = 100$. Фрагменты этого множества приведены в таблице 1. На первом этапе классификации было решено сформировать три кластерных образования, начальные положения геометрических центров которых расположены в точках $e_0^1 = (255000, 75,5)$; $e_0^2 = (750000, 425,5)$; $e_0^3 = (1350000, 850,5)$, координаты их выражены в натуральных единицах населения и площади. Радиусы же в безразмерных величинах оказались равными $r_0^1 = 7,2$; $r_0^2 = 20,2$; $r_0^3 = 12,7$. Процесс пошагового преобразования кластерных сфероидов с параметром точности $\varepsilon = 0,5$ представлен в таблице 2 и визуализирован на рисунке 2. На нем пунктирными линиями очерчены промежуточные положения кластеров, а сплошными – их финальные состояния и размеры.

Таблица 1 – Фрагмент списка городов РФ с указанием численности населения и территории

№	Город	Население, X	Y	Площадь, км^2, W	V
1	Новосибирск	1 602 915	41,92	506,67	46,14
2	Уфа	1 115 560	32,15	707	60,46
3	Орск	230 414	14,41	621,33	54,34
4	Казань	1 231 878	34,48	314,16	53,82
5	Волжский	326 055	16,33	229,12	26,30
6	Омск	1 178 391	33,41	566,9	50,45
7	Самара	1 169 719	33,24	541,382	48,62
8	Ростов-на Дону	1 125 299	32,35	348,5	34,83
…	…	…	…	…	…
137	Дербент	123 162	12,26	69,63	14,90
…	…	…	…	…	…
284	Кинешма	185 368	18,73	78,67	15,71
…	…	…	…	…	…
306	Энгельс	208 604	20,12	84,47	16,12

Таблица 2 – Этапы преобразования кластеров

№	Центр кластера [население, площадь]	Радиус r
1		
1	[220698, 102,5]	7,2
2	[227734, 108,1]	7,2
2		
1	[567468, 334,4]	10,2
2	[461127, 316,11]	8,2
3	[442760, 305,50]	7,2
4	[439437, 301,9]	8,2
5	[443997, 293,5]	6,2
3		
1	[10559716, 788,7]	29,7
2	[978274, 557,8]	13,7
3	[1024047, 214,5]	11,7
4	[1073611, 485,8]	10,7
5	[1117677, 480,0]	11,7
6	[1117677, 480,0]	10,7

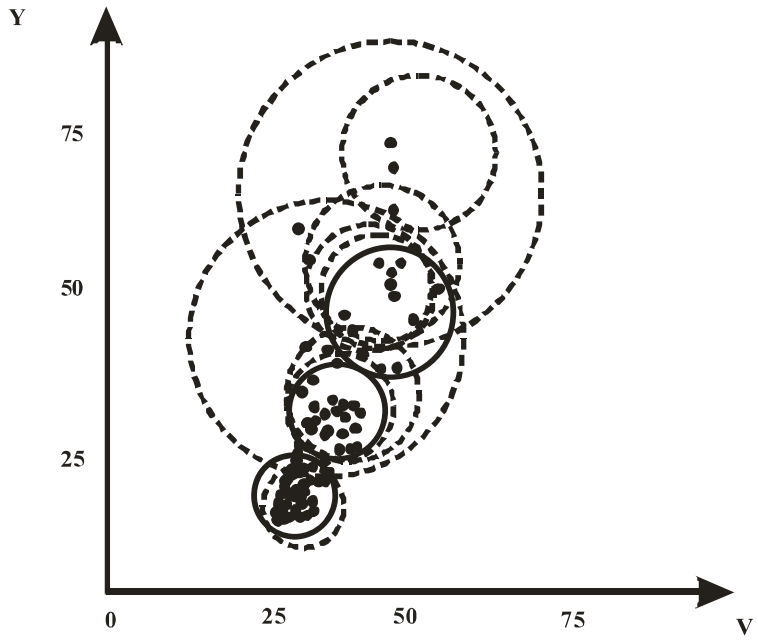


Рисунок 2 – Процесс коррекции кластерных сфероидов

Из данных таблицы видно, что финальная позиция первого кластера была достигнута после двух корректирующих воздействий, не затронувших величину радиуса. Это объясняется тем, что плотность группировки в зоне кластера S1 наиболее высока (малые города больше похожи друг на друга, чем мегаполисы). Работа со вторым и третьим кластером потребовала пять и шесть корректирующих воздействий соответственно. Они коснулись не только положений геометрических центров, но и радиусов. Это оказалось обязательным для предотвращения пересечения кластеров. Интересным результатом, по мнению авторов, представляющим самостоятельное значение, является значительное (почти в три раза) уменьшение радиуса кластерного сфероида S^3 по сравнению с начальным и заметное смещение положения его геометрического центра в сторону меньших значений признака X-населения. Это указывает на то, что города с населением 800–900 тысяч жителей ближе по структуре к мегаполисам, чем к городам, население которых насчитывает 500–600 тысяч человек, а также подтверждает тезис о самоорганизации алгоритма.

На рисунке 2 ясно видно, что часть объектов исходного массива данных не вошла в состав ни одного из трех созданных кластеров, и так как поместить их туда не представляется возможным, то они остаются вне классификации. Отдельно для таких объектов «изгоев» алгоритм был запущен еще раз, что привело к созданию еще трех кластеров: S^4, S^5, S^6, что визуализировано на рисунке 3. Их радиусы равны $r_0^4 = 11,2$; $r_0^5 = 7,2$; $r_0^6 = 1,9$. Число элементов, вошедших в состав этих кластеров, меньше чем в трех первых, что нисколько не снижает значимость полученных результатов. В условиях потоков пополнения и гибели, а также естественных изменений объектов (рост или убыль населения, изменения границ городской черты и т.п.), положение кластеров и мощность множеств элементов в их составе будет меняться, что и делает актуальной идею динамической кластеризации.

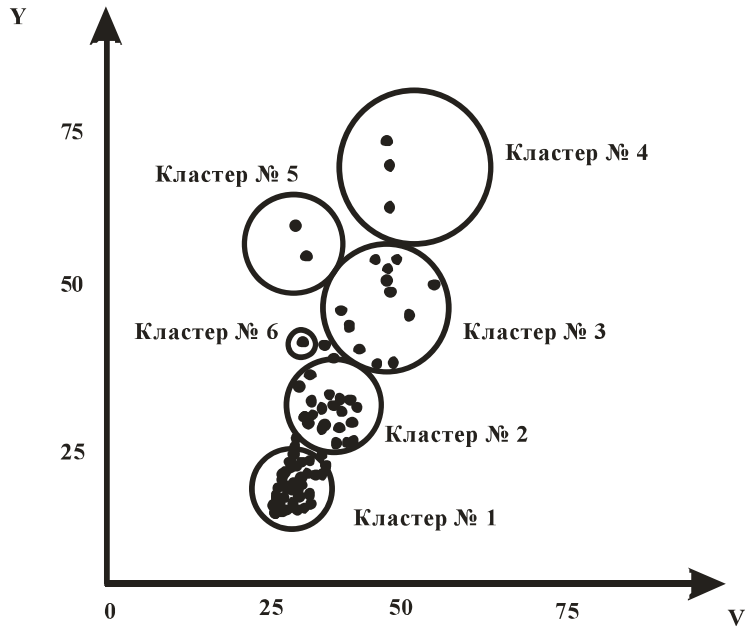


Рисунок 3 – Этапы и повторный результат расчетов, выполненные в соответствии с разработанным алгоритмом

Авторы располагают сведениями о применении описанного здесь алгоритма для обработки больших данных. По заказу одной территориальной телекоммуникационной компании была предпринята успешная попытка классификации ее клиентов. В качестве классификационных признаков, которых набралось более 100, были предложены различные услуги и сервисы, предоставляемые пользователям и некоторые данные пользователей. Общий объем клиентской базы компании составлял ≈1000000 единиц, имея незначительные колебания нерегулярного характера по времени. Обработка этого массива позволила выделить базы 9 категорий пользователей, различающихся по характеру их предпочтений. Результаты были использованы для создания портрета «типичного клиента» и выработки рациональной ценовой политики по комплексу предлагаемых опций. К сожалению, мы не можем опубликовать эти материалы, поскольку они являются собственностью компании, где проводилось исследование. Руководство компании не сочло возможным дать разрешение на их публикацию, считая эти материалы коммерческой тайной.

Выводы:

1. Разработан алгоритм, позволяющий осуществлять кластеризацию массивов «больших данных» в динамике их изменений.
2. Как показала практика, алгоритм эффективен для обработки данных, характеризуемых большим числом количественно измеримых признаков.
3. Алгоритм обладает свойствами самоорганизуемости и самообучаемости, т.е. в автоматическом режиме осуществляет коррекцию положения и размеров кластеров в факторном пространстве с учетом пространственных ограничений.
4. В работе представлен и проиллюстрирован демонстрационный пример с подробной интерпретацией результатов.

Библиографический список

1. Аль-Хашеди А. А. Экспертиза методов, используемых в различных задачах распознавания образов / А. А. Аль-Хашеди, Е. А. Печеный, Н. К. Нуриев // Вестник технологического университета. – 2017. – Т. 20, № 1. – С. 125–127.
2. Воронцов К. В. Алгоритмы кластеризации и многомерного шкалирования / К. В. Воронцов. – Москва : МГУ, 2007. – 234 с.
3. Гарднер М. Математические головоломки и развлечения / М. Гарднер. – Москва : Мир, 1971. – 510 с.
4. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл. – Москва : Статистика, 1977. – 128 с.
5. Ершов К. С. Анализ и классификация алгоритмов кластеризации / К. С. Ершов, Т. Н. Романова // Новые информационные технологии в автоматизированных системах. – 2016. – № 19. – С. 274–279.
6. Журавлёв Ю. И. Распознавание. Классификация. Прогноз. Математические методы и их применение / Ю. И. Журавлёв. – Москва : Наука, 1989. – 302 с.
7. Игнатьев Н. А. Кластерный анализ данных и выбор объектов-эталонов в задачах распознавания с учителем / Н. А. Игнатьев // Вычислительные технологии. – 2015. – Т. 20, № 6. – С. 36–45.

8. Колмогоров А. Н. Элементы теории функций и функционального анализа / А. Н. Колмогоров, С. В. Фомин. – Москва : Физматлит, 2004. – 572 с.

9. Низаметдинов Ш. У. Анализ данных : учебное пособие / Ш. У. Низметдинов, А. П. Румянцев. – Москва : МИФИ, 2012. – 286 с.

10. Нуриев Н. К. Математическое моделирование эволюции кластерных образований / Н. К. Нуриев, А. А. Аль-Хашеди, Е. А. Печеный // Современные наукоемкие технологии. – 2018. – № 8. – С. 110–116.

11. Coates A. Learning Feature Representations with K-means / A. Coates, A. Y. Ng. // Neural Networks: Tricks of the Trade. – 2012. – P. 561–580.

12. Berkhim P. Survey of Clustering Data Mining Techniques / P. Berkhim // Accue Software. – 2002. – 160 p.

13. Griffin G. Learning and using taxonomies for fast visual categorization / G. Griffin, P. Perona // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2008. – P. 1–8.

14. Ilango M. A survey of grid based clustring algorithms / M. Ilango, V. Mohan // Intern. J. of Eng. Sci. and Technology. – 2010. – Vol. 2 (8). – P. 3441–3446.

15. Jain A. K. Data Clustering: A review / A. K. Jain, M. N. Marty, P. J. Flynn // ACM Computing Surveys, 1999. – Vol. 31, № 3. – P. 264–323.

16. Kagan J. Clustering Lange and High Dimensional Data / J. Kagan, C. Nicholas, M. Teloulle. – Режим доступа: http://www.csee.umtc.edu/nicholas/clustering/tutorial.pdf, свободный. – Заглавие с экрана. – Яз. рус.

17. Kumar J. Parallel using large data sets / J. Kumar, R. T. Mills, F. M. Hoffman, W. W. Hargrove // Procedia Computer Science. – 2011. – № 4. – P. 1602–1611.

18. Shaukat K. Dengue Fever in Perspective of Clustering Algorithms / K. Shaukat, N. Masood, A. B. Shafaat, K. Jabbar, H. Shabbir et al. // Data Mining Genomics Proteomics. – 2015. – Vol. 6, № 176. DOI:10.4172/2153-0602.1000176.

19. Ximing Lv. Research on P2P Network Loan Risk Evaluation Based on Generalized DEA Model and R-Type Clustering Analysis under the Background of Big Data / Lv. Ximing, Lan Zhou, Xiaona Guo // Financial Risk Management. – 2017. – Vol. 6, № 2. – P. 163–190.

20. Yin J. A. Developmental approach to structural self-organization in reservoir computing / J. Yin, Y. Meng, Y. Jin // IEEE Transactions on Autonomous Mental Development. – 2012. – № 4 (4). – P. 273–289.

References

1. Al-Khashedi A. A., Pechenyy E. A., Nuriev N. K. Ekspertiza metodov, ispolzuyemykh v razlichnykh zadachakh raspoznavaniya obrazov [Examination of methods used in various pattern recognition tasks]. *Vestnik tekhnologicheskogo universiteta* [Bulletin of the University of Technology], 2017, vol. 20, no. 1, pp. 125–127.

2. Vorontsov K. V. *Algoritmy klasterizatsii i mnogomernogo shkalirovaniya* [Algorithms for clustering and multidimensional scaling]. Moscow, Moscow State University Publ., 2007. 234 p.

3. Gardner M. *Matematicheskiye golovolomki i razvlecheniya* [Math puzzles and fun]. Moscow, Mir Publ., 1971. 510 p.

4. Durant B., Odell P. *Klasternyy analiz* [Cluster analysis]. Moscow, Statistika Publ., 1977. 128 p.

5. Ershov K. S., Romanova T. N. Analiz i klassifikatsiya algoritmov klasterizatsii [Analysis and classification of clustering algorithms]. *Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems], 2016, no. 19, pp. 274–279.

6. Zhuravlev Yu. I. *Raspoznavaniye. Klassifikatsiya. Prognoz. Matematicheskiye metody i ikh primeneniye* [Recognition. Classification. Forecast. Mathematical methods and their application]. Moscow, Nauka Publ., 1989. 302 p.

7. Ignatiev N. A. Klasternyy analiz dannykh i vybor obektov-etalonov v zadachakh raspoznavaniya s uchitelem [Cluster data analysis and selection of objects-standards in tasks of recognition with a teacher]. *Vychislitel'nyye tekhnologii* [Computational Technologies], 2015, vol. 20, no. 6, pp. 36–45.

8. Kolmagorov A. N., Fomin S. V. *Elementy teorii funktsiy i funktsionalnogo analiza* [Elements of the theory of functions and functional analysis]. Moscow, Fizmatlit Publ., 2004. 572 p.

9. Nizametdinov W. W., Rumyantsev A. P. Analiz dannykh : uchebnoye posobiye [Data Analysis : Tutorial]. Moscow, Moscow Engineering Physics Institute, 2012. 286 p.

10. Nuriev N. K., Al-Hashedi A. A., Pechenyy E. A. Matematicheskoye modelirovaniye evolyutsii klasternykh obrazovaniy [Mathematical modeling of the evolution of cluster formations]. *Sovremennyye naukoyemkiye tekhnologii* [Modern High Technologies], 2018, no. 8, pp. 110–116.

11. Coates A., Ng. A. Y. Learning Feature Representations with K-means. *Neural Networks: Tricks of the Trade*, 2012, pp. 561–580.

12. Berkhim P. Survey of Clustering Data Mining Techniques. *Accue Software*, 2002. 160 p.

13. Griffin G., Perona P. Learning and using taxonomies for fast visual categorization. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

14. Ilango M., Mohan V. A survey of grid based clustring algorithms. *Intern. J. of Eng. Sci. and Technology*, 2010, vol. 2 (8), pp. 3441–3446.

15. Jain A. K., Marty M. N., Flynn P. J. Data Clustering: A review. *ACM Computing Surveys*, 1999, vol. 31, no. 3, pp. 264–323.

16. Kagan J., Nicholas C., Teloulle M. *Clustering Lange and High Dimensional Data*. Available at: http://www.csee.umtc.edu/nicholas/clustering/tutorial.pdf.

17. Kumar J., Mills R. T., Hoffman F. M., Hargrove W. W. Parallel using large data sets. *Procedia Computer Science*, 2011, no. 4, pp. 1602–1611.

18. Shaukat K., Masood N., Shafaat A. B., Jabbar K., Shabbir H. et al. Dengue Fever in Perspective of Clustering Algorithms. *Data Mining Genomics Proteomics*, 2015, vol. 6, no. 176. DOI:10.4172/2153-0602.1000176.

19. Ximing Lv. Zhou Lan, Guo Xiaona. Research on P2P Network Loan Risk Evaluation Based on Generalized DEA Model and R-Type Clustering Analysis under the Background of Big Data. *Financial Risk Management*, 2017, vol. 6, no. 2, pp. 163–190.

20. Yin J. A., Meng Y., Jin Y. Developmental approach to structural self-organization in reservoir computing. *IEEE Transactions on Autonomous Mental Development*, 2012, no. 4 (4), pp. 273–289.

DOI 10.21672/2074-1707.2020.49.4.020-032
УДК 004:614

НЕЙРОСЕТЕВАЯ ТЕХНОЛОГИЯ ОБНАРУЖЕНИЯ АНОМАЛЬНОГО СЕТЕВОГО ТРАФИКА

Статья поступила в редакцию 05.12.2019, в окончательном варианте – 26.02.2020.

Частикова Вера Аркадьевна, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
кандидат технических наук, доцент, e-mail: chastikova_va@mail.ru

Жерлицын Сергей Анатольевич, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
студент, e-mail: kpytooooo@gmail.com

Воля Яна Игоревна, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
студент, e-mail: volya_y@mail.ru

Сотников Владимир Владимирович, Кубанский государственный технологический университет, 350072, Российская Федерация, г. Краснодар, ул. Московская, 2,
студент, e-mail: bubert9@mail.ru

Рассмотрены существующие методы анализа сетевого трафика, указаны их возможности и ограничения. Продемонстрирована актуальность решаемой задачи. Обоснована целесообразность использования нейросетевого подхода к обнаружению аномалий сетевого трафика. Исследована эффективность использования алгоритмов роевого интеллекта применительно к задаче обучения нейронных сетей, выявлены особенности данных алгоритмов. Реализована объектно-ориентированная библиотека для выявления сетевых атак с использованием нейросети с архитектурой многослойного персептрона. На данном этапе исследования для обучения нейросети и оценки качества распознавания трафика был применен датасет KDD Cup 1999 Data. Описаны преимущества и недостатки реализованного решения. Представлен способ устранения распространенного недостатка датасетов, связанного с несбалансированностью обучающих данных. Описаны используемые технологии: архитектура нейронной сети, алгоритм обучения, способ уменьшения размерности обрабатываемых данных. На втором этапе был использован набор данных CSE-CIC-IDS2018. Предложена нейросетевая модель, построенная на базе архитектуры LSTM и эмбеддинговой сетей. Для обучения разработанной системы предложено применение алгоритма Adam, основанного на градиентном спуске. На основе использования названных алгоритмов, моделей и технологий был реализован, а затем и протестирован программный комплекс для обнаружения сетевых атак.

Ключевые слова: нейронная сеть, сетевая атака, многослойный перцептрон, роевой интеллект, LSTM-сеть, эмбеддинговая сеть, Focal Loss, алгоритм Adam

NEURAL NETWORK TECHNOLOGY FOR DETECTING ANOMALOUS NETWORK TRAFFIC

The article was received by the editorial board on 05.12.2019, in the final version – 26.02.2020.

Chastikova Vera A., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
Cand. Sci. (Engineering), Associate Professor, e-mail: chastikova_va@mail.ru

Zherlitsyn Sergey A., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
student, e-mail: kpytooooo@gmail.com

Volya Yana I., Kuban State Technological University, 2 Moskovskaya St., Krasnodar, 350072, Russian Federation,
student, e-mail: volya_y@mail.ru