УДК 004.21

Р.Г. Асадуллаев, А.Н. Афонин, М.А. Ситникова

R.G. ASADULLAEV, A.N. Afonin, M.A. Sitnikova

**АЛГОРИТМ КЛАСТЕРИЗАЦИИ НЕЙРОНАЛЬНОЙ АКТИВНОСТИ МОЗГА ПРИ РАСПОЗНАВАНИИ ПАТТЕРНОВ ДВИЖЕНИЙ ВЕРХНИХ КОНЕЧНОСТЕЙ**

**ALGORITHM OF CLUSTERING OF NEURONAL ACTIVITY OF THE BRAIN WHEN RECOGNIZING THE PATTERNS OF TOP MOUNT MOVEMENTS**

*Статья посвящена проблеме анализа данных нейрональной активности мозга при распознавании паттернов движений верхних конечностей с целью построения системы управления бионическим протезом оснащенным механическими приводами. Разрабатывается модифицированный алгоритм укрупненной кластеризации, в основе которого заложен алгоритм DBSCAN. Предлагается при первоначальной укрупненной кластеризации формировать множество шум-кластеров из точек, не попавших в выделенные кластеры вместо формирования единственной категории “шум”. Это позволит при поступлении новых данных проводить оперативный анализ точек, не попавших в выделенные кластеры на предмет формирования новых кластеров из шум-кластеров при достижении заданной плотности. В результате разработан алгоритм непрерывной кластеризации, который в процессе формирования центроидов кластеров может проводить дополнительное разбиение уже сформированных кластеров на более мелкие в силу их неправильной формы. При этом алгоритм постоянно оценивает вновь поступающие данные на предмет соответствия выявленным кластерам или шум-кластерам. В результате становится возможным формирование новых кластеров при достижении шум-кластерами минимальной плотности.*

*Ключевые слова: анализ многомерных данных, анализ нейрональной активности мозга, кластерный анализ, DBSCAN.*

*The article is devoted to the problem of analyzing the data of brain neuronal activity in recognition of patterns of upper limb movements with the aim of constructing a control system for a bionic prosthesis equipped with mechanical drives. A modified algorithm for large-scale clustering is developed, based on the DBSCAN algorithm. It is suggested that at the initial large clustering to form a lot of noise clusters from points that did not fall into the allocated clusters instead of forming a single category of "noise". This will allow, upon receipt of new data, to conduct an operative analysis of points that have not fallen into the allocated clusters for the formation of new clusters from noise clusters upon reaching a given density. As a result, an algorithm for continuous clustering has been developed that, in the process of forming centroids of clusters, can further partition the already formed clusters into smaller clusters because of their irregular shape. At the same time, the algorithm constantly evaluates the incoming data for compliance with the identified clusters or noise clusters. As a result, it becomes possible to form new clusters when the noise density reaches clusters.*

*Keywords: analysis of multidimensional data, analysis of brain neuronal activity, cluster analysis, DBSCAN.*

Бионические протезы, оснащенные механическими приводами, являются на сегодняшний день наиболее эффективными устройствами для реабилитации инвалидов с ампутированными конечностями. Существующие исследования, направленные на разработку и создание бионических протезов, сосредоточены в двух основных направлениях: удешевление самого протеза и улучшение системы управления [1]. Перспективным направлением получения данных о движениях конечностей является динамическая визуализация и непрерывный анализ в режиме реального времени изображений нейрональной активности посредством технологии нейрообратной связи с помощью спектроскопии в околоинфракрасном диапазоне (функциональной fNIRS-томографии). fNIRS-томография – это техника оптической визуализации, в которой измеряемое относительное количество поглощенного с определенной длиной волны инфракрасного света может указывать на относительные концентрации окси- и дезоксигемоглобина в изучаемой области головного мозга, которые могут быть рассчитаны путем применения модифицированного закона Ламберта-Бера [2, 3, 4].

Кластеризация данных позволяет сформировать относительно однородные группы объектов с учетом определенного сходства на основании имеющейся выборки данных. Отсюда вытекает основная характеристика кластеров, заключающаяся во внутренней однородности и внешней изолированности. В зависимости от решаемой задачи, отличительной особенностью кластерного анализа от классификации является начальная неопределенность классов и их свойств, а также в некоторых задачах их количество. Данные параметры задаются исследователем при проведении кластерного анализа [5, 6, 7].

При проведении кластерного анализа необходимо задать степень сходства между объектами анализа, которая может быть сформулирована посредством вычисления значения расстояния между данными. При этом могут быть использованы следующие метрики: евклидово расстояние, представляющее собой геометрическое расстояние в многомерном пространстве; квадрат евклидова расстояния, позволяющий придать больший вес отдаленным друг от друга объектам; манхэттенское расстояние (среднее разностей по координатам); расстояние Чебышева (применяется когда два объекта имеют различные значения одной из координат и нужно определить их как различные); степенное расстояние (при необходимости увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются). Метрика сходства объектов анализа задается исследователем в зависимости от задачи кластеризации. При этом при решении задачи кластеризации могут быть использованы несколько метрик.

Наиболее распространенными алгоритмами кластеризации иерархического типа являются восходящие алгоритмы ближнего, среднего и дальнего соседа. Работа данных алгоритмов основана на том, что на начальном этапе каждый элемент выборки представляется как отдельный кластер, после чего производится итеративное объединение наиболее близких друг к друг кластеров [8]. Наиболее известными алгоритмами неиерархической итеративной кластеризации является метод k-means (k-средних) и его разновидность метод k-medoids [9, 10, 11].

Плотностные алгоритмы формируют кластеры как совокупность кучно расположенных объектов, то есть кластер формируется на основании заданного радиуса, в котором должно находиться некоторое минимальное количество объектов. Представителем плотностных алгоритмов является DBSCAN [12, 13]. Статистические алгоритмы кластеризации основаны на предположении, что кластеры описываются некоторым семейством вероятностных распределений.

Существует ряд задач, требующих особого подхода при проведении кластерного анализа многомерных данных. Алгоритм кластеризации данных исходной выборки должен удовлетворять следующим требованиям: способен работать с большим объемом многомерных данных; не требует указания искомого количества кластеров; не требует указания максимальной величины кластера; не чувствителен к шумам; не требует обучения; легко поддается модификации и вариации параметров; модифицирует множество кластеров с учетом поступления новых данных.

В результате анализа алгоритмов кластеризации иерархических и неиерархических типов на предмет удовлетворения указанным требованиям к функционированию выявлено, что алгоритмы иерархического типа на начальном этапе требуют указания максимального размера кластера (максимально возможного количества элементов), что ограничивает их применение при отсутствии данной информации. Итеративные алгоритмы неиерархического типа не могут быть использованы в связи с необходимостью указания количества искомых кластеров и чувствительностью к шумам. Неиерархические статистические алгоритмы плохо справляются с анализом больших данных. Наиболее подходящими к предъявляемым требованиям оказались плотностные алгоритмы. Таким образом, реализация алгоритма кластеризации данных, удовлетворяющего заданным требованиям, будет осуществлена на базе алгоритма DBSCAN с учетом дополнительной модификации.

В процессе реализации алгоритма непрерывной кластеризации многомерных данных необходима предварительная процедура укрупненной кластеризации имеющихся данных. С целью учета специфики алгоритма осуществлена модификация алгоритма DBSCAN (рисунок 1). Основная модификация заключается в формировании множества шум-кластеров из точек, не попавших в выделенные кластеры вместо формирования единственной категории “шум”. При этом распределение точек шума по шум-кластерам осуществляется в зависимости от их координат. Это позволит в дальнейшем при поступлении новых данных проводить оперативный анализ точек, не попавших в выделенные кластеры на предмет формирования новых кластеров из шум-кластеров.



Рисунок 1 - Модифицированный алгоритм DBSCAN

Таким образом, работа алгоритма (рисунок 1) заключается в следующих действиях:

1. Ввод исходных данных X = $\{x\_{1}, x\_{2},…,x\_{n}\}$, где $x\_{i}$ – точка в многомерном пространстве. и определение параметров $ε$ – значение радиуса окрестности и $MinPts$ *–* минимальное количество точек-соседей для формирования кластера. Параметры $ε$ и $MinPts$ выбираются в зависимости от предметной области и интенсивности поступления новых данных. Формирование массивов кластеров (K) размером pK и шум-кластеров (NK) размером pNK, содержащих сведения о количестве точек в кластере.

2. Если имеются не обработанные данные, то переход к шагу 3, в противном случае вывод данных, представляющих собой сведения о массивах K и NK.

3. Из загруженного массива данных случайным образом выбирается элемент для анализа, представляющий собой точку в многомерном пространстве. Затем осуществляется проверка окрестности точки на наличие соседей, представляющих собой точки, расстояние до которых меньше заданного радиуса окрестности $p(x,y)\leq ε$, где $p(x,y)$ представляет собой евклидово расстояние, рассчитываемое относительно $n$ точек попавших в радиус заданной окрестности $ε$.

4. Если количество точек – соседей в окрестности больше либо ровно MinPts, то создается новый кластер inc(pK), где pK – текущее количество кластеров. В противном случае, когда количество соседствующих точек меньше MinPts, принимается решение о создании шум-кластера inc(pNK), где pNK – текущее количество шум-кластеров. Тем же образом, что и для полноценного кластера, который описан выше. Точка, относительно которой проверялась окрестность записывается в кластер либо шум-кластер и помечается, как пройденная. Остальные точки из рассматриваемой окрестности помещаются во временную категорию. Если в окрестности анализируемого объекта не оказалось точек-соседей, шум-кластер создается из одного элемента – текущей точки.

5. Если во временной категории имеются точки, то случайным образом осуществляется переход к любой точке из категории, для которой также осуществляется проверка окрестности на наличие точек соседей. После чего реализуется запись текущей точки в текущий кластер либо шум-кластер и помечается как пройденная. Если в проверяемую окрестность из временной категории попали новые точки, ранее не учтенные во временной категории, то они также записываются в текущую временную категорию и реализуется переход к шагу 5. В противном случае осуществляется просто переход к шагу 5. Действие повторяется до тех пор, пока во временной категории не закончатся точки, т.е. пока алгоритм не дойдет до границы кластера. Таким образом формируется текущий кластер. После того, как во временной категории закончились точки, осуществляется переход к шагу 2, то есть проверяется условие наличия не обработанных точек в массиве входных данных и при выполнении условия случайным образом выбирается одна из них, и процедура проверки окрестности повторяется для нее.

На рисунке 2 представлен алгоритм непрерывной кластеризации многомерных данных, удовлетворяющий сформированным выше требованиям и в основе которого лежит разработанный модифицированный алгоритм кластеризации данных DBSCAN.

Алгоритм (рисунок 2) формирует кластеры различных размеров и особые шум-кластеры. Специфика работы алгоритма позволяет оперативно модифицировать множество кластеров посредством непрерывного сбора и анализа вновь поступающих данных. Работа алгоритма заключается в выполнении следующих итераций:

1. Ввод исходных данных, представляющих собой имеющиеся наблюдения о системе.

2. Предварительная укрупненная кластеризация данных модифицированным алгоритмом DBSCAN (рисунок 1), результатом работы которого являются pK кластеров и pNK шум-кластеров.

3. Для каждого *Ki*-го кластера, где $i=\left[1,pK\right] $ определяется центроид Ci в зависимости от количества элементов в кластере.

Выявленные центроиды определяют области наибольшего скопления элементов выборки. Кроме того, в результате работы модифицированного плотностного алгоритма DBSCAN возможно получение кластеров неправильной формы – вытянутых, лентовидных, спиральных и т.д. В этом случае весовой центр кластера определить не удастся, для чего алгоритмом предусмотрена локальная кластеризация с меньшим размером окрестности, направленная на распределение объектов выявленного кластера на более мелкие. Таким образом, в результате локальной кластеризации i-й кластер преобразуется locK новых более мелких кластеров, для которых также применяется процедура определения центральных точек.

4. После того, как для каждого кластера определен центроид осуществляется непрерывная обработка вновь поступающих данных до тех пор, пока их обработка будет актуальна для исследователя. Для каждого вновь полученного элемента проводится проверка на принадлежность к выделенным кластерам согласно тому же подходу, который используется при разбиении выборки объектов на группы – при помощи проверки окрестности нового объекта. Если кластер определен, то новые данные записываются в соответствующий кластер, в противном случае осуществляется проверка на принадлежность к имеющимся шум кластерам.



Рисунок 2 – Алгоритм непрерывной кластеризации многомерных данных

5. В том случае, если шум-кластер не определен, создается новый шум-кластер, состоящий из этой точки. Если же элемент отнесен к одному из шум-кластеров, то для элементов шум-кластера осуществляется проверка плотности окрестности с учетом начального значения минимума плотности. Решение о создании нового кластера из шум-кластера принимается при выполнении условия, говорящего о том, что число точек в окрестности центроида $p(x,y)\leq ε$ больше либо ровно $MinPts$.

Предложенный алгоритм непрерывной кластеризации данных представляет собой комбинацию из модифицированного алгоритма DBSCAN и набора оригинальных процедур, учитывающих вновь поступающие данные. Относительно алгоритмов кластеризации многомерных данных, рассмотренных в ходе исследования, предложенный алгоритм не требует задания ожидаемого количества кластеров и позволяет перестраивать структуру распределения кластеров в многомерном пространстве с учетом новых данных. В состав алгоритма входят оригинальные процедуры, позволяющие декомпозировать кластеры неправильной формы на более мелкие, что позволяет более детально проводить дальнейшую классификацию данных и принимать более обоснованные решения, учитывающие специфику распределения данных и индивидуального характера поведения каждого кластера в многомерном пространстве. Так при изучении нейрональной активности мозга для распознавания паттернов движений верхних конечностей с целью построения системы управления бионическим протезом оснащенным механическими приводами, разработанный алгоритм позволит формировать кластеры, отвечающие за отдельные движения. При этом в алгоритме реализована возможность постоянного мониторинга новых данных и перераспределения кластерного пространства.

Исследование выполнено при финансовой поддержке РФФИ и Администрации Белгородской областив рамках научного проекта 18-48-310028 р\_а.

**СПИСОК ЛИТЕРАТУРА**

1. Афонин А.Н. Разработка и реализация макета бионического протеза кисти руки / А.Н. Афонин, А.Ю. Алейников, А.Р.Гладышев, А.В. Попова // Робототехника и техническая кибернетика, 2016, №3(12). - С. 68-71.

2. Ferrari M., Quaresima V. A brief review on the history of human functional near- infrared spectroscopy (fNIRS) development and fields of application // Neuroimage. - 2012. - V. 63. – Pp. 921–935.

3. Scholkmann F., Kleiser S., Metz A.J., Zimmermann R., MataPavia J., Wolf U. et al. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology // Neuroimage. - 2014. - V. 85. – Pp. 6–27.

4. Ситникова М.А., Нюрк Г.Х. Функциональная оптическая томография: надежный метод измерения мозговой активации в процессе решения различных математических задач // Современные проблемы науки и образования. – 2016, №3.

5. Перелыгин, А.А., 2015. Кластеризация Многомерных Данных: Методы, Алгоритмы, Программы. Вестник Алтайского государственного педагогического университета: естественные и точные науки, 25: 24-31.

6. Ester, M.A., Kriegel, H.-P., Sander, J. and Xu, X., 1998. Clustering for Mining in Large Spatial Databases. Issue on Data Mining, KI-Journal, ScienTec Publishing, 1: 1-7.

7. [Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavav, P., 1998. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In Proc. ACM SIGMOD’98 Int. Conf. on Management of Data, Seattle, Washington, pp: 94-105.](http://www.cs.cornell.edu/johannes/papers/1998/sigmod1998-clique.pdf)

8. Alam, S., Dobbie, G. and Rehman, S.U., 2015. Analysis of particle swarm optimization based hierarchical data clustering approaches. Journal Swarm and Evolutionary Computation, 25: 36–51.

9. Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Journal Pattern Recognition Letters, 31: 651 – 666.

10. Rejito, J., Retantyo, W., Hartati, S. and Harjoko, A. 2012. Optimization CBIR using K-Means Clustering for Image Database. International Journal of Computer Science and Information Technologies, 3(4): 4789-4793.

11. Niknam, T., Amiri, B., 2010. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Journal Applied Soft Computing, Elsevier, 10: 183-197.

12. [Kailing, K., Kriegel, H.-P. and Kröger, P., 2014. Density-Connected Subspace Clustering for High-Dimensional Data. In Proceedings of the 2004 SIAM International Conference on Data Mining, pp: 246-256.](http://www.siam.org/proceedings/datamining/2004/dm04_023kailingk.pdf)

13. Тханг, В.В., Пантюхин, Д.В. и Галушкин, А.И., 2015. Гибридный алгоритм кластеризации FastDBSCAN. Труды МФТИ, 3(7): 77-81.

**Асадуллаев Рустам Геннадьевич**

ФГАОУ ВО «Белгородский государственный технологический университет», г. Белгород

К.т.н., доцент кафедры прикладной информатики и информационных технологий

Тел.: +7(4722)30-12-94

E-mail: asadullaev@bsu.edu.ru

**Афонин Андрей Николаевич**

ФГАОУ ВО «Белгородский государственный технологический университет», г. Белгород

Д.т.н., профессор кафедры материаловедения и нанотехнологий

Тел.: +7(4722)30-18-28

E-mail: afonin@bsu.edu.ru

**Ситникова Мария Александровна**

ФГАОУ ВО «Белгородский государственный технологический университет», г. Белгород

К.псих.н., доцент кафедры психологии

Тел.: +7(4722)30-28-65

E-mail: furmanchuk@bsu.edu.ru