ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ, АНАЛИЗ ДАННЫХ, МАШИННОЕ ОБУЧЕНИЕ

УДК 004.9 ISSN 1995-5499
DOI: https://doi.org/
Поступила в редакцию 11.01.2022
Подписана в печать 22.04.2022

ВЯЗКИЙ ГРАВИТАЦИОННЫЙ АЛГОРИТМ КЛАСТЕРИЗАЦИИ НЕТОЧНЫХ ДАННЫХ

© 2022 П. А. Головинский✉

Воронежский государственный технический университет, ул. 20-летия Октября, 84, 394006 Воронеж, Российская Федерация

Аннотация. Кластеризация является одной из базовых задач машинного обучения, наряду с распознаванием образов, классификацией и прогнозированием. Особенно существенна роль кластеризации в анализе больших данных, работа с которыми может быть эффективной только с использованием компьютерных технологий. При этом, задача автоматического разбиения на кластеры с учетом погрешностей исходных данных не получила однозначного решения и требует поиска более адекватных подходов, включающих автоматическое определение числа кластеров. В работе предложен новый метод кластеризации данных, основанный на модификации гравитационного алгоритма, использующего аналогию с формированием звездных кластеров за счет притяжения масс в соответствии с законом всемирного тяготения. При применении такого подхода к кластеризации данных реальные физические массы заменяются точками в многомерном пространстве данных, а движение этих точек с учетом их притяжения приводит к формированию кластеров. Недостатком такого способа является проявление эффектов инерции, которые могут затруднять процесс завершения кластеризации и приводить к выбросу ускоренных частиц из кластера на стадии его формирования. Для исключения таких нежелательных событий в работе используется модель динамики вязкого движения частиц, представляющих данные, и естественное ограничение размеров кластеров за счет отталкивания частиц. Силы отталкивания частиц взяты в виде обменного взаимодействия Паули для фермионов при гауссовом распределении плотностей погрешностей. Записаны основные уравнения, описывающие работу представленной модификации гравитационного алгоритма. На численном примере продемонстрированы особенности и преимущества вязкого гравитационного алгоритма в сравнении с методом k-средних и основанном на плотностях методом DBSCAN, включая автоматическую остановку процедуры при завершении процесса кластеризации. Полученные результаты позволяют проводить слепую кластеризацию больших данных и допускают обобщение на решение задач многомерной оптимизации.

Ключевые слова: кластеризация данных, неточные данные, гравитационный алгоритм, вязкость, отталкивание Паули.

ВВЕДЕНИЕ

При исследовании систем, содержащих очень большое количество элементов, например в анализе больших данных, важнейшим этапом является первичная классификация без учителя, которая в математическом смысле эквивалентна кластеризации [1]. По ее результатам происходит разделение элементов на небольшое количество классов, внутри каждого из которых все элементы характеризуются относительной однородностью.

✉ Головинский Павел Абрамович
e-mail: golovinski@bk.ru

Кластеризация позволяет перейти от анализа и описания отдельных представителей класса к описанию типичных особенностей элемента из данного кластера. Это дает далее возможность прогнозировать поведение групп однотипных элементов на основе временной статистики и распознавать принадлежность нового элемента к тому или иному классу.

Развитие технологии Big Data сделало необходимым обеспечить надежное решение задачи кластеризации компьютерными методами [2] без предварительного анализа имеющейся информации человеком. Для алгоритмов кластеризации важна их адаптивность к изменяющемуся количеству объектов, устойчивость против выбросов в векторах признаков или неизвестных тяжелых хвостов распределений характеристик, последовательность работы с постоянным обновлением для потоковой передачи данных без необходимости повторного запуска всего алгоритма и простота вычислений.

1. СОСТОЯНИЕ ПРОБЛЕМЫ И ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

Для решения задач кластеризации развит впечатляющий набор иерархических (hiererhical), разделительных (partitional) и байесовских (Bayesian) алгоритмов [3–5], начиная от классического алгоритма k-средних и самоорганизующихся карт Кохонена, и кончая алгоритмами гравитационной кластеризации. Общим недостатком большинства этих методов является необходимость начального задания полного числа кластеров. Если перейти к размытым, нечетким данным, то инструменты нечеткой логики позволяют решить задачу кластеризации вместе с определением числа кластеров [6]. Число кластеров определяется и в подходах, основанных на введение свойства плотности элементов при образовании кластеров [7, 8].

В имеющихся алгоритмах кластеризации неточных данных сохраняются проблемы, касающиеся релевантности, быстродействия и универсальности в применении к данным с различной топологией кластеров. При этом обычно предполагается, что имеющиеся в распоряжении данные соответствуют выборке из ансамбля с некоторым, чаще всего гауссовым, распределением случайных данных. Одним из активно развиваемых методов кластеризации с естественным механизмом формированием кластеров является алгоритм на основе гравитационной физической аналогии [9, 10], который привлекает наглядностью и простотой расчета. В нем данным сопоставляют отдельные частицы в многомерном пространстве, испытывающие взаимное притяжение подобное притяжению единичных масс. Гравитационный алгоритм нашел многочисленные применения в задачах управления и принятия решений [11–15]. Кроме того, гравитационный алгоритм позволяет решать сложные задачи глобальной оптимизации [16–20].

Несмотря на существенное удобство использования алгоритмов гравитационной кластеризации и оптимизации в распределенных системах, ему присущи определенные недостатки. Так, уравнения движения Ньютона с консервативными силами, применяемые для описания процесса кластеризации, уже в исходной физической формулировке приводят к таким нежелательным эффектам, как возможный выброс частиц из области кластеризации за счет инерции [21], поскольку движение частиц продолжается даже после прекращения действия сил. Кроме того, алгоритм не имеет естественного критерия завершения, и нужно вводить специальные механизмы останова, используя, например, конечные размеры элементов, отражающие неточность задания исходных данных.

В целом, полностью удовлетворительное решение задачи кластеризации еще не найдено, и усилия в этом направлении продолжаются. Настоящая работа описывает алгоритм кластеризации неточных данных, в котором используется аналогия с агрегированием частиц конечного размера в вязкой жидкости за счет их притяжения и слипания в кластеры при наличии отталкивания на малых расстояниях. Для устранения эффектов инерции, движение фиктивных частиц многомерного пространства происходит в условной вязкой

среде, где их скорость пропорциональна приложенной силе. Это представление позволяет также понизить порядок решаемой системы дифференциальных уравнений и ускорить ее численное решение в процессе компьютерного моделирования.

2. МЕТОДЫ И МАТЕРИАЛЫ

В качестве прототипа метода кластеризации данных мы возьмем модель физической кластеризации частиц под действием взаимных сил притяжения частиц в вязкой среде [22], пренебрегая инерцией. Движение таких частиц описывается системой уравнений первого порядка

$$\frac{d\mathbf{x}_j}{dt} = \lambda \mathbf{F}_j, \qquad (1)$$

где $\mathbf{x}_j$ — n-мерный вектор, задающий положение j-й частицы, $\mathbf{F}_j$ — суммарная сила, действующая на частицу, λ — коэффициент, отражающий величину вязкости среды и регулирующий скорость образования кластеров. При парном взаимодействии N частиц, на частицу с номером j действует сила

$$\mathbf{F}_j = \sum_{i \neq j}^{N} \mathbf{f}_{ji}. \qquad (2)$$

Если парные силы являются центральными и зависят только от расстояния между частицами, то

$$\mathbf{f}_{ji} = \mathbf{x}_{ji} d\left(\left\|\mathbf{x}_{ji}\right\|\right), \qquad (3)$$

где $\left\|\mathbf{x}_{ji}\right\|$ — евклидово расстояние, $d\left(\left\|\mathbf{x}_{ji}\right\|\right)$ — скалярная функция.

В стандартном гравитационном алгоритме (GSA) $\mathbf{x}_j$ представляет собой вектор данных, а процесс их кластеризации описывается уравнениями Ньютона, согласно которым $m_j \ddot{\mathbf{x}}_j = \mathbf{F}_j$, т. е. в отличие от предлагаемого нами метода, — дифференциальными уравнениями второго порядка. Обычно, в гравитационном алгоритме предполагается $d\left(\left\|\mathbf{x}_{ji}\right\|\right) \sim 1/\left(\left\|\mathbf{x}_{ji}\right\|^p + \varepsilon\right)$, $p = 2$ [23–25]. Тем самым, в GSA в действительности используется нефизический гравитационный закон для силы [26], поскольку обычные силы гравитации $(p = 3)$ убывают с расстоянием слишком значительно, чтобы обеспечить достаточно быструю кластеризацию системы.

Разработаны различные версии, уточняющие и дополняющие алгоритм GSA [27–31]. Отметим, что в гравитационной кластеризации можно увидеть черты, общие с самоорганизующимися картами Кохонена. Основное отличие состоит в том, что в картах Кохонена имеются пробные векторы, которые потом перемещаются в сторону того или иного кластера, а в гравитационной кластеризации обычно все частицы стягиваются в кластеры за счет взаимного притяжения.

Предлагаемый нами гравитационный алгоритм с вязкостью описывает диссипативную динамику и исключает выброс частиц из кластеров за счет высвобождения гравитационной энергии образующих кластер частиц. Алгоритм может автоматически останавливаться, если в него включить конечные размеры частиц или их отталкивание на малых расстояниях. Тогда при соприкосновении частиц их дальнейшее сближение прекращается. После присоединения к одному из кластеров последней частицы состав кластеров полностью определен, а их центры можно определить как среднее значение для элементов кластера. Данный алгоритм определяет фиксированное количество кластеров и положение их центров за конечное время.

В предлагаемом алгоритме мы определим способ задания силы отталкивания частиц на близком расстоянии, пользуясь квантово-механической аналогией. Как известно, фермионы с одинаковыми спинами обладают квантовым отталкиванием Паули, величина которого пропорциональна перекрытию плотностей электронных распределений [32]. Эта физическая модель позволяет учесть неопределенность задания элементов кластеризации. Будем считать, что неопределенность задания элемента с номером j описывается плотностью распределения $\rho_j(\mathbf{r}_j)$. Введем парную силу отталкивания пропорциональную свертке плотностей распределения пары данных в виде

$$\mathbf{f}_{ji}^{P} = -\frac{\mathbf{x}_{ji}}{\left|\mathbf{x}_{ji}\right|} C \int \rho_j(\mathbf{x} - \mathbf{x}_j) \rho_i(\mathbf{x} - \mathbf{x}_i)\, d\mathbf{x}. \qquad (4)$$

Величина гиперпараметра C выбирается таким образом, чтобы компенсировать дальнодействующую силу притяжения на расстоянии, соответствующем соприкосновению частиц. В качестве такого расстояния естественно выбрать сумму радиусов распределений для выбранной пары данных. Радиус распределения, при его гауссовой форме, определяется дисперсией. Автоматический останов сближения частиц i и j без учета влияния других частиц происходит при выполнении условия

$$\mathbf{f}_{ji} + \mathbf{f}_{ji} = 0. \qquad (5)$$

При наличии многих частиц останов происходит при равенстве нулю всех сил и прекращении движения частицы. Этот момент рассматривается как присоединение частицы к кластеру и фиксируется заданием соответствующего элемента матрицы смежности в графе кластеризации.

Алгоритм вязкой гравитационной кластеризации отражен в Листинге.

Листинг

1: **input** данные $\boldsymbol{X}$ и погрешности $\boldsymbol{\theta}$;

2: **preprocessing** масштабирование исходных данных $\boldsymbol{X} \to \boldsymbol{Y}(0)$, $\boldsymbol{\theta} \to \boldsymbol{\sigma}$;

3: **initialize** начальное время $t = 0$, шаг по времени dt и время кластеризации T;

4: **while** $(t < T)$ **do**

5: Вычисление массива сил $\boldsymbol{F}$, действующих на частицы;

6: Вычисление нового положения частиц $\boldsymbol{Y}(t + dt) = \boldsymbol{Y}(t) + dt \cdot \boldsymbol{F}$, шага dt и времени $t = t + dt$;

7: **end while**

8: Выделение и разметка кластеров $\boldsymbol{X} \to \boldsymbol{A}$ для близких по $\boldsymbol{\sigma}$ точек $\boldsymbol{Y}(T)$;

9: **return** метки кластеров $\boldsymbol{A}$

Таким образом, рассматриваемый нами алгоритм кластеризации данных отличается от GSA использованием вязкости и отталкивания в форме Паули. Исходные данные могут обладать произвольной анизотропной неопределенностью, что может быть учтено соответствующим заданием функций $\rho_j(\mathbf{x})$.

3. РЕЗУЛЬТАТЫ ЧИСЛЕННОГО МОДЕЛИРОВАНИЯ

При выборе модели распределения погрешностей данных мы воспользуемся n–мерным гауссовым распределением векторной величины $\mathbf{x}$ со средним значением $\boldsymbol{\mu}$ и матрицей ковариаций $\boldsymbol{\Sigma}$ в виде

$$\rho(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{(2\pi)^{-p/2}}{\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \qquad (6)$$

Здесь $\mathbf{x}^T$ есть транспонированный вектор $\mathbf{x}$, а $\left|\boldsymbol{\Sigma}\right|$ есть детерминант матрицы $\boldsymbol{\Sigma}$. Для двух многомерных нормальных распределений с плотностями функций $\rho(\mathbf{x};\mathbf{a},\mathbf{A})$ и $\rho(\mathbf{x};\mathbf{b},\mathbf{B})$ свертка представляет собой вновь нормальное распределение $\rho(\mathbf{x};\mathbf{a}-\mathbf{b},\mathbf{A}+\mathbf{B})$, поскольку

$$\int \rho(\mathbf{x};\mathbf{a},\mathbf{A})\rho(\mathbf{x};\mathbf{b},\mathbf{B})\,d\mathbf{x} = \\ = \frac{(2\pi)^{-p/2}}{\left|\mathbf{A}+\mathbf{B}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{a}-\mathbf{b})^T(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{a}-\mathbf{b})}. \qquad (7)$$

Результат симметричен относительно параметров двух распределений, и формула (7) удобна для расчета силы отталкивания при сближении центров распределений.

Для сравнения работы алгоритма использовался метод k-средних и основанная на плотности пространственная кластеризация для приложений с шумами (DBSCAN) [33, 34]. Алгоритм кластеризации DBSCAN основан на исследовании плотности точек в пространстве данных. Он группирует вместе точки, которые тесно расположены, помечая как выбросы те из них, которые находятся одиноко в областях с малой плотностью, т.е. точки, ближайшие соседи которых лежат далеко.

На рис. 1 представлен результат кластеризации случайного двумерного набора данных из 200 точек методом k-средних. Результат кластеризации соответствуют зрительной оценке кластеров.

Несмотря на то, что DBSCAN является одним из наиболее используемых алгоритмов кластеризации, и часто упоминается в научной литературе, его применение может приводить к довольно парадоксальным результатам. Применение метода DBSCAN к тому же набору данных приводит к совсем иной

кластеризации, представленной на рис. 2, где основная часть данных образует один кластер, в то время как в другом кластере содержится только одна точка, соответствующая выбросу данных. Такой результат наглядно показывает, что алгоритм DBSCAN хуже разделяет кластеры с данными, концентрирующимися вблизи отдельных центров, по сравнению с методом k-средних, несмотря на то, что он способен проводить кластеризацию данных с более сложной топологией. Только при более значительном пространственном разделении кластеров метод DBSCAN в нашем примере определяет еще один наблюдаемый кластер из большого числа точек.

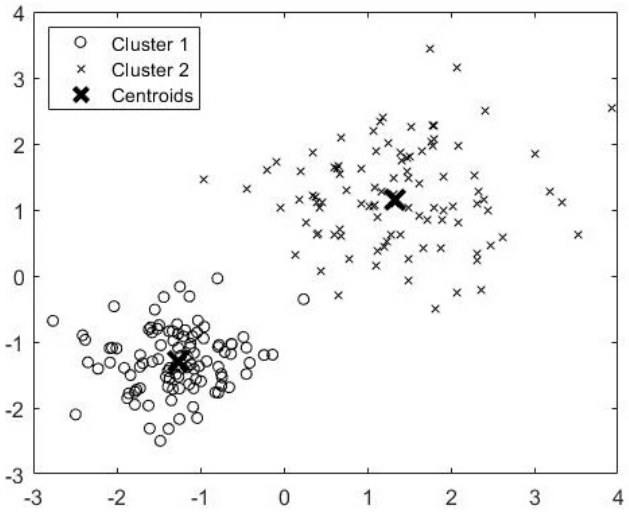

Рис. 1. Кластеризация данных методом k-средних

[Fig. 1. Clustering data using the k-means method]

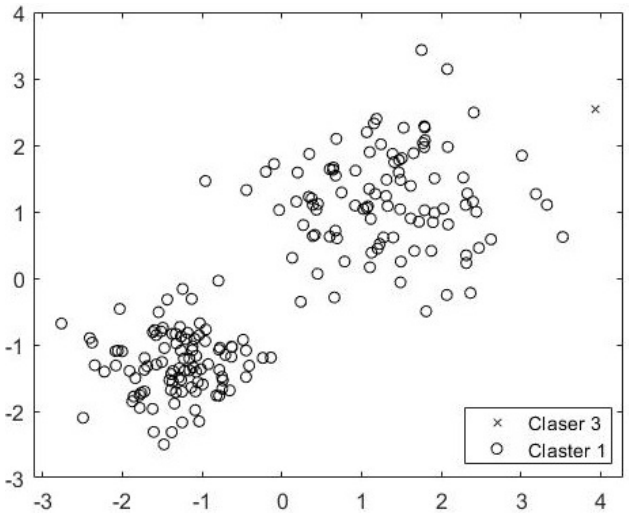

Рис. 2. Кластеризация данных методом DBSCAN

[Fig. 2. Clustering data using the DBSCAN method]

Применение к тому же набору данных алгоритма гравитационного агрегирования с вязкостью и отталкиванием в форме Паули приводит к выделению трех кластеров. На рис. 3 представлены результаты агрегирования данных за счет притяжения точек. В исходные значения внесены гауссовы распределения погрешностей

$$\rho(x,y;a,b)=\exp\left(1-\frac{(x-a)^2+(y-b)^2}{\sigma^2}\right) \quad (8)$$

с параметром $\sigma=0,01$. Масштабирование условного времени t выбрано в соответствии со значением $\lambda=1$.

Процесс стягивания в плотные агрегаты резко замедляется после формирования первичных кластеров. Это автоматически решает задачу определения числа кластеров, являющейся проблемой во многих других алгоритмах кластеризации. Размеры области агрегированных точек определяет точность определения соответствующего центра кластера. Первый кластер полностью совпадает с кластером, полученным методом k-средних. Второй кластер отличается от результата метода k-средних только одной точкой, представляющий выброс и образующий отдельный третий кластер.

На рис. 4 показан результат разбиения данных на кластеры вязким гравитационным алгоритмом. Первые два кластера воспроизводят кластеры, полученные в методе k-сред-

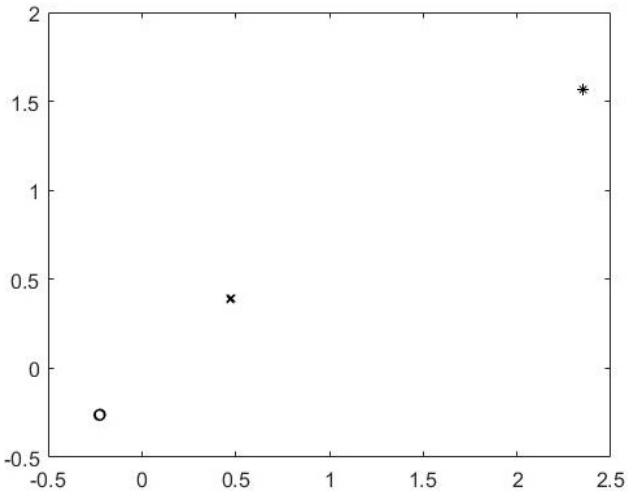

Рис. 3. Гравитационное агрегирование данных

[Fig. 3. Gravitational data aggregation]

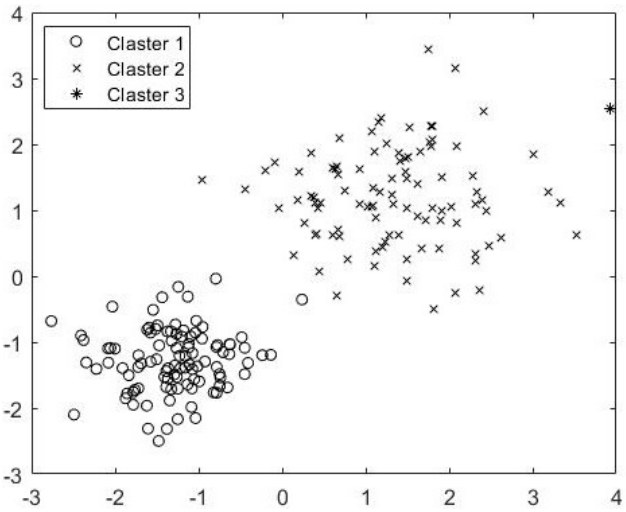

Рис. 4. Кластеризация данных гравитационным алгоритмом
[Fig. 4. Clustering of data by the gravitational algorithm]

них, а третий кластер содержит только одну точку (*), являющуюся выбросом, который был зафиксирован как отдельный кластер методом DBSCAN.

Принадлежность новых данных к тому или иному кластеру в нашем методе определяется определением максимальной силы, действующей со стороны каждого кластера на новую пробную частицу. Для кластера из одной точки этот способ определения принадлежности для выбранного нами закона силы эквивалентен выбору ближайшего центра кластера в евклидовой метрике.

Если разместить центры агрегатов, получившихся в результате гравитационного стягивания, в центре масс соответствующих кластеров, то процесс кластеризации можно продолжить в рамках того же алгоритма и с новыми данными, не исключая формирование новых кластеров.

ЗАКЛЮЧЕНИЕ

Предложенный в работе вариант гравитационного алгоритма кластеризации отличается от известных версий двумя важными особенностями. В качестве физического прототипа выбрано притяжение частиц не в свободном пространстве, а в среде с большой вязкостью. В этом случае агрегирование частиц происходит в соответствии с динамическими уравнениями для их скоростей, пропорциональных интегральной действующей на соответствующую частицу силе.

Для моделирования неточности задания данных использовался потенциал отталкивания, выбранный по аналогии с отталкиванием Паули, характерным для ферми-частиц в квантовой теории. Такой подход позволяет получать агрегирование точек в кластеры с определенной точностью, задаваемой точностью исходных данных и обеспечивает автоматический останов алгоритма за счет выравнивания сил притяжения и отталкивания при достаточном сближении частиц. При использовании номинальных данных неопределенность в их задании можно рассматривать на основании нечеткой математики и описывать с помощью соответствующих гауссовских функций принадлежности. После этого можно вновь применять предложенный в работе вариант гравитационного алгоритма кластеризации.

Вязкий гравитационный алгоритм может быть распространен на задачи поиска многомерной оптимизации с помощью подхода, основанного на рое частиц с переменной массой. В гравитационном алгоритме оптимизации, массы частиц, представляющих данные, увеличиваются с ростом величины целевой функции. Благодаря этому, начальный случайный рой частиц стягивается в область максимальных значений целевой функции, обеспечивая нахождение экстремума. Отметим, что в обычном варианте гравитационный алгоритм поиска был опробован не только на тестовых функциях, но и применен к важной задаче определения наилучшего расположения тепловыделяющих стержней в ядерном реакторе [35]. Применение вязкого гравитационного алгоритма к многомерным задачам оптимизации режимов работы энергетических объектов представляет интерес для дальнейших исследований.

КОНФЛИКТ ИНТЕРЕСОВ

Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Evans, R.* Clustering for classification / R. Evans, B. Pfahringer and G. Holmes // 2011 7th International Conference on Information Technology in Asia. – Kuching, Sarawak, 2011. – P. 1–8.

2. *Shirkhorshidi, A. S.* Big Data clustering: A review / A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, T. Herawan // In: Murgante B. et al. (eds) Computational Science and Its Applications – ICCSA 2014. Lecture Notes in Computer Science. – Springer, Cham. – 2014. – V. 8583. – P. 707–720.

3. *Ge, M.* Big Data for Internet of Things: A Survey / M. Ge, H. Bangui, B. Buhnova // Future Generation Computer Systems. – 2018. –V. 87. – P. 601–614.

4. *Xu, R.* Survey of clustering algorithms / Rui Xu, D. Wunsch // IEEE Transactions on Neural Networks. – 2005. – V. 16, No 3. – P. 645–678.

5. *Xu, D.* A Comprehensive survey of clustering algorithms / D. Xu, Y. Tian // Ann. Data. Sci. – 2015. – V. 2. – P. 165–193.

6. *Jain, A. K.* Data clustering: 50 years beyond K-means / A. K. Jain // Pattern Recognition Letters. – 2010. – V. 31(8). – P. 651–666.

7. *Lemke, O.* Density-based cluster algorithms for the identification of core sets / O. Lemke, B. G. Keller // J. Chem. Phys. – 2016. – V. 145(16). – P. 164104(14).

8. *Lee, K. M.* Density and frequency-aware cluster identification for spatio-temporal sequence data / K. M. Lee, S. Y. Lee, et al // Wireless Pers. Commun. – 2017. – V. 93. – P. 47–65.

9. *Wright, W. E.* Gravitational clustering / W. E. Wright // Pattern Recognition. – 1977. – V. 9. – P. 151–166.

10. *Gorbonos, D.* Similarities between insect swarms and isothermal globular clusters / D. Gorbonos, K. van der Vaart, M. Sinhuber, J. G. Puckett, A. M. Reynolds, N. T. Ouellette, N. S. Gov. // Phys. Rev. Research. – 2020. – V. 2. – P. 013271(5).

11. *Fazliana Abdul Kadir A.* An improved gravitational search algorithm for optimal placement and sizing of renewable distributed generation units in a distribution system for power quality enhancement / A. Fazliana Abdul Kadir, A. Mohamed, H. Shareef, A. Asrul Ibrahim, T. Khatib, W. Elmenreich // Journal of Renewable and Sustainable Energy. – 2014. – V. 6(3). – P. 033112(17).

12. *Mahdad, B.* Interactive gravitational search algorithm and pattern search algorithms for practical dynamic economic dispatch / B. Mahdad, K. Srairi // International Transactions on Electrical Energy Systems. – 2014. – V. 25(10). – P. 2289–2309.

13. *Kou, Z.* Association rule mining using chaotic gravitational search algorithm for discovering relations between manufacturing system capabilities and product features / Z. Kou // Concurrent Engineering. – 2019. – V. 27(3). – P. 213–232.

14. *Huang, M.-L.* Combining a gravitational search algorithm, particle swarm optimization, and fuzzy rules to improve the classification performance of a feed-forward neural network / M.-L. Huang, Y.-C. Chou // Computer Methods and Programs in Biomedicine. – 2019. – V. 180. – P. – 105016(12).

15. *Siddique, N.* Applications of gravitational search algorithm in engineering / N. Siddique, H. Adeli // Journal of Civil Engineering and Management. – 2016. – V. 22(8). – P. 981–990.

16. *Ali, A. F.* Direct gravitational search algorithm for global optimisation problems / A. F. Ali, M. A. Tawhid // East Asian Journal on Applied Mathematics. – 2016. – V. 6(03). – P. 290–313.

17. *Koay, Y. Y.* An adaptive gravitational search algorithm for global optimization / Y. Y. Koay, J. D. Tan, C. W. Lim, S. P. Koh, S. K. Tiong, K. Ali // Indonesian Journal of Electrical Engineering and Computer Science. – 2019. – V. 16(2). – P. 724–729.

18. *Zhang, A.* A hybrid genetic algorithm and gravitational search algorithm for global optimization / A. Zhang, G. Sun, Z. Wang, Y. Yao // Neural Network World. – 2015. –V. 25(1). – P. 53–73

19. *Rashedi, E.* A comprehensive survey on gravitational search algorithm. / E. Rashedi, E. Rashedi, H. Nezamabadi-pour // Swarm and Evolutionary Computation. – 2018. – V. 41. – P. 141–158.

20. *Xiaobing, Y.* An improved gravitational search algorithm for global optimization /

Y. Xiaobing, Y. Xianrui, C. Hong // J. Intell. Fuzzy Syst. – 2019. – V. 37. – P. 50395047.

21. *Vasile, M.* Incremental planning of multi-gravity assist trajectories / M. Vasile, J. M. R. Martin, L. Masi, E. Minisci, R. Epenoy, V. Martinot; J. F. Baig // Acta Astronautica. – 2015. – V. 115. – P. 407–421.

22. *Binder, P.* Gravitational clustering: A simple, robust and adaptive approach for distributed networks / P. Binder, M. Muma, A. M. Zoubir // Signal Processing. – 2018. – V. 149. – P. 36–48.

23. *Rashedi, E.* GSA: A gravitational search algorithm / E. Rashedi, H. Nezamabadi-pour, S. Saryazdi // Information Sciences. – 2009. – V. 179(13). – P. 2232–2248.

24. *Sabri, N. M.* A review of gravitational search algorithm / N. M. Sabri, M. Puth, M. R. Mahmood // Int. J. Advance. Soft. Comp. Appl. – 2013. – V. 5, No. 3. – P. 1-39.

25. *Bala, I.* Gravitational search algorithm: A state-of-the-art review. / I. Bala, A. Yadav // In: Yadav N., Yadav A., Bansal J., Deep K., Kim J. (eds) Harmony Search and Nature Inspired Optimization Algorithms. Advances in Intelligent Systems and Computing.– Springer, Singapore. – 2019. – V. 741. – P. 27–37.

26. *Gauci, M.* Why 'CSA: a gravitational search algorithm' is not genuinely based on the law of gravity / M. Gauci, T. J. Dodd, R. Groß // Nat. Comput. – 2012. – V. 11. – P. 719–720.

27. *Alswaitti, M.* Optimized gravitational-based data clustering algorithm / M. Alswaitti, M. K. Ishak, N. A. M. Isa // Engineering Applications of Artificial Intelligence. – 2018. – V. 73. – P. 126–148.

28. *Gomez, J.* The Parameter-less Randomized Gravitational Clustering algorithm with online clusters' structure characterization / J. Gomez, E. Leon, O. Nasraoui, F. Giraldo // Prog. Artif. Intell. – 2014. – V. 2. – P. 217–236.

29. *Han, X.* A novel data clustering algorithm based on modified gravitational search algorithm / X. Han, L. Quan, X. Xiong, M. Almeter, J. Xiang, Y. Lan // Engineering Applications of Artificial Intelligence. – 2017. – V. 61. – P. 1–7.

30. *Rashedi, E.* A comprehensive survey on gravitational search algorithm / E. Rashedi, E. Rashedi, H. Nezamabadi-pour // Swarm and Evolutionary Computation. – 2018. – V. 41. – P. 141–158.

31. *Mustafa, H. M. J.* An improved adaptive memetic differential evolution optimization algorithms for data clustering problems / H. M. J. Mustafa, M. Ayob, M. Z. A. Nazri, G. Kendall // PLOS ONE. – 2019. – V. 14(5). – P. e0216906(28).

32. *Slamet, M.* Rigorous and unifying physical interpretation of the exchange potential and energy in the local-density approximation / M. Slamet, V. Sahni // Phys. Rev. B. – 1992. – V. 45(8). – P. 4013–4019

33. *Sander, J.* Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications / J. Sander, M. Ester, HP. Kriegel, X. Xu // Data Mining and Knowledge Discovery. – 1998. – V. 2. – P. 169–194.

34. *Rodriguez, A.* Clustering by fast search and find of density peaks / A. Rodriguez, A. Laio // Science. – 2014. – V. 344(6191). – P. 1492–1496.

35. *Mahmoudi, S. M.* A novel optimization method, Gravitational Search Algorithm (GSA), for PWR core optimization / S. M. Mahmoudi, M. Aghaie, M. Bahonar, N. Poursalehi // Annals of Nuclear Energy. – 2016. – V. 95. – P. 23–34.

Головинский Павел Абрамович – д-р физ.-мат. наук, профессор кафедры инноватики и строительной физики им. И. С. Суровцева Воронежского государственного технического университета.
E-mail: golovinski@bk.ru
ORCID iD: https://orcid.org/0000-0002-7527-0297

DOI: https://doi.org/
Received 11.01.2022
Accepted 22.04.2022

ISSN 1995-5499

VISCOUS GRAVITATIONAL ALGORITHM FOR CLUSTERING INACURATE DATA

© 2022 P. A. Golovinski✉

Voronezh State Technical University
84, 20-letiya Oktyabrya Street, 394006 Voronezh, Russian Federation

Annotation. Clustering is one of the basic problems of machine learning, along with pattern recognition, classification and forecasting. The role of clustering is especially important in the analysis of Big Data, work with which can only be carried out using computer technologies. At the same time, the problem of automatic partitioning into clusters, taking into account the errors of the initial data, has not up to now an unambiguous solution and requires a search for more adequate approaches, including automatic determination of the number of clusters. The paper proposes a new method for data clustering, based on a modification of the gravitational algorithm, which uses an analogy with the formation of stellar clusters due to the attraction of masses in accordance with the law of universal gravitation. When applying this approach to data clustering, real physical masses are replaced by points in a multidimensional data space, and the motion of these points, taking into account their attraction, leads to the formation of clusters. The disadvantage of this method is the manifestation of the effects of inertia, which can hinder the clustering process and lead to the ejection of accelerated particles from the cluster at the stage of its formation. To exclude such phenomena, we use a model of the dynamics of viscous motion of particles representing the data and the natural limitation of the cluster size due to the repulsion of particles. When simulating the repulsive force of particles, the interaction in the Pauli form was taken for fermions with the same spins and the Gaussian distribution of the error density. The basic equations describing the steps of the presented modification of the gravitational algorithm are written. A numerical example demonstrates the features and advantages of the viscous gravity algorithm in comparison with the k-means method and the density-based DBSCAN method, including automatic termination of the procedure when the main clustering process is completed. The results obtained allow for blind clustering of Big Data, and can be generalized to solving multidimensional optimization problems.
Keywords: data clustering, imprecise data, gravity algorithm, viscosity, Pauli repulsion.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Evans R., Pfahringer B. and Holmes G.* (2011) Clustering for classification. *2011 7th International Conference on Information Technology in Asia. Kuching, Sarawak*. P. 1–8.

2. *Shirkhorshidi A. S., Aghabozorgi S., Wah T. Y. and Herawan T.* (2014) Big Data clustering: A review. *In: Murgante B. et al. (eds) Computational Science and Its Applications – ICCSA 2014. Lecture Notes in Computer Science*. Springer, Cham. 8583. P. 707–720.

3. *Ge M., Bangui H. and Buhnova B.* (2018) Big Data for Internet of Things: A Survey. *Future Generation Computer Systems*. 87. P. 601–614.

4. *Xu Rui and Wunsch D.* (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. 16(3). P. 645–678.

5. *Xu D. and Tian Y.* (2015) A Comprehensive survey of clustering algorithms. *Ann. Data. Sci*. 2. P. 165–193.

✉ Golovinski Pavel A.
e-mail: golovinski@bk.ru

6. *Jain A. K.* (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 31(8). P. 651–666.

7. *Lemke O. and Keller B. G.* (2016) Density-based cluster algorithms for the identification of core sets. *J. Chem. Phys*. 145(16). P. 164104(14).

8. *Lee K. M., Lee S. Y. [et al]* (2017) Density and frequency-aware cluster identification for spatio-temporal sequence data. *Wireless Pers. Commun*. 93. P. 47–65.

9. *Wright W. E.* (1977) Gravitational clustering. *Pattern Recognition*. 9. P. 151–166.

10. *Gorbonos D., van der Vaart K., Sinhuber M., Puckett J. G., Reynolds A. M., Ouellette N. T. and Gov N. S.* (2020) Similarities between insect swarms and isothermal globular clusters. *Phys. Rev. Research*. 2. P. 013271(5).

11. *Fazliana Abdul Kadir A., Mohamed A., Shareef H., Asrul Ibrahim A., Khatib T. and Elmenreich W.* (2014) An improved gravitational search algorithm for optimal placement and sizing of renewable distributed generation units in a distribution system for power quality enhancement. *Journal of Renewable and Sustainable Energy*. 6(3). P. 033112(17).

12. *Mahdad B. and Srairi K.* (2014) Interactive gravitational search algorithm and pattern search algorithms for practical dynamic economic dispatch. *International Transactions on Electrical Energy Systems*. 25(10). P. 2289–2309.

13. *Kou Z.* (2019) Association rule mining using chaotic gravitational search algorithm for discovering relations between manufacturing system capabilities and product features. *Concurrent Engineering*. 27(3). P. 213– 232.

14. *Huang M.-L. and Chou Y.-C.* (2019) Combining a gravitational search algorithm, particle swarm optimization, and fuzzy rules to improve the classification performance of a feed-forward neural network. *Computer Methods and Programs in Biomedicine*. 180. P. – 105016(12).

15. *Siddique N. and Adeli H.* (2016) Applications of gravitational search algorithm in engineering / *Journal of Civil Engineering and Management*. 22(8). P. 981–990.

16. *Ali A. F. and Tawhid M. A.* (2016) Direct gravitational search algorithm for global optimisation problems. *East Asian Journal on Applied Mathematics*. 6(03). P. 290–313.

17. *Koay Y. Y., Tan J. D., Lim C. W., Koh S. P., Tiong S. K. and Ali K.* (2019) An adaptive gravitational search algorithm for global optimization. *Indonesian Journal of Electrical Engineering and Computer Science*. 16(2). P. 724–729.

18. *Zhang A., Sun G., Wang Z. and Yao Y.* (2015) A hybrid genetic algorithm and gravitational search algorithm for global optimization. *Neural Network World*. 25(1). P. 53–73

19. *Rashedi E., Rashedi E. and Nezamabadi-pour H.* (2018) A comprehensive survey on gravitational search algorithm. *Swarm and Evolutionary Computation*. 41. P. 141–158.

20. *Xiaobing Y., Xianrui Y. and Hong C.* (2019) An improved gravitational search algorithm for global optimization. *J. Intell. Fuzzy Syst.* 37. P. 50395047.

21. *Vasile M., Martin J. M. R., Masi L., Minisci E., Epenoy R., Martinot V. and Baig J. F.* (2015) Incremental planning of multi-gravity assist trajectories. *Acta Astronautica*. 115. P. 407–421.

22. *Binder P., Muma M. and Zoubir A. M.* (2018) Gravitational clustering: A simple, robust and adaptive approach for distributed networks. *Signal Processing*. 149. P. 36–48.

23. *Rashedi E., Nezamabadi-pour H. and Saryazdi S.* (2009) GSA: A gravitational search algorithm. *Information Sciences*. 179(13). P. 2232–2248.

24. *Sabri N. M., Puth M. and Mahmood M.* (2013) R. A review of gravitational search algorithm. *Int. J. Advance. Soft. Comp. Appl.* 5. (3). P. 1–39.

25. *Bala I. and Yadav A.* (2019) Gravitational search algorithm: A state-of-the-art review. In: Yadav N., Yadav A., Bansal J., Deep K., Kim J. (eds) *Harmony Search and Nature Inspired Optimization Algorithms. Advances in Intelligent Systems and Computing*. Springer, Singapore. 741. P. 27–37.

26. *Gauci M., Dodd T. J. and Groß R.* (2012) Why 'CSA: a gravitational search algorithm' is not genuinely based on the law of gravity. *Nat. Comput*. 11. P. 719–720.

27. *Alswaitti M., Ishak M. K. and Isa N. A. M.* (2018) Optimized gravitational-based data clustering algorithm. *Engineering Applications of Artificial Intelligence*. 73. P. 126–148.

28. *Gomez J., Leon E., Nasraoui O. and Giraldo F.* (2014) The Parameter-less Randomized Gravitational Clustering algorithm with online clusters' structure characterization. *Prog. Artif. Intell*. 2. P. 217–236.

29. *Han X., Quan L., Xiong X., Almeter M., Xiang J. and Lan Y.* (2017) A novel data clustering algorithm based on modified gravitational search algorithm. *Engineering Applications of Artificial Intelligence*. 61. P. 1–7.

30. *Rashedi E. and Nezamabadi-pour H.* (2018) A comprehensive survey on gravitational search algorithm. *Swarm and Evolutionary Computation*. 41. P. 141–158.

31. *Mustafa H. M. J., Ayob M., Nazri M. Z. A. and Kendall G.* (2019) An improved adaptive memetic differential evolution optimization algorithms for data clustering problems. *PLOS ONE*. 4(5). P. e0216906(28).

32. *Slamet M. and Sahni V.* (1992) Rigorous and unifying physical interpretation of the exchange potential and energy in the local-density approximation. *Phys. Rev. B*. 45(8). P. 4013–4019

33. *Sander J., Ester M., Kriegel HP. and Xu X.* (1998) Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*. 2. P. 169–194.

34. *Rodriguez A. and Laio A.* (2014) Clustering by fast search and find of density peaks. *Science*. 344(6191). P. 1492–1496.

35. *Mahmoudi S. M., Aghaie M., Bahonar M. and Poursalehi N.* (2016) A novel optimization method, Gravitational Search Algorithm (GSA), for PWR core optimization. *Annals of Nuclear Energy*. 95. P. 23–34.

Golovinski Pavel A. — Dr. Phys.-Math. Sci., Professor of the Department of Innovation and Building Physics named after I. S. Surovtsev, Voronezh State Technical University.
E-mail: golovinski@bk.ru
ORCID iD: https://orcid.org/0000-0002-7527-0297