Синюк В. Г., канд. техн. наук, доц.,
Поляков В. М., канд. техн. наук, доц.,
Анищенко А. И., студент
Белгородский государственный технологический университет им. В.Г. Шухова

ОБ ОДНОМ ГИБРИДНОМ АЛГОРИТМЕ НЕЧЁТКОЙ КЛАСТЕРИЗАЦИИ*

vgsinuk@mail.ru

Нечёткая кластеризация является важнейшей проблемой и предметом активных исследований в различных областях. Нечёткий алгоритм С-среднего один из самых популярных методов нечёткой кластеризации (FCM). Однако FCM чувствителен к инициализации и легко попадает в локальные оптимумы. Во многих задачах оптимизации используется Particle Swarm Optimization (PSO) – инструмент глобальной стохастической оптимизации. В данной статье предлагается гибридный нечёткий метод оптимизации, базируемый на FCM и нечёткой PSO (FPSO), который использует достоинств обоих алгоритмов. Эксперименты показывают, что предлагаемый метод эффективен и может показывать хорошие результаты.

Ключевые слова: *FCM, FPSO, нечёткая кластеризация, гибридный алгоритм.*

Введение

Кластеризация – это процесс сопоставления объектов данных с набором непересекающихся групп называемых кластерами так, чтобы объекты в каждом кластере были более похожи друг на друга, чем на объекты других кластеров.

С-среднего – это один из самых популярных жёстких алгоритмов кластеризации, который распределяет объекты данных на С кластеров (К-количество кластеров).

Нечёткие алгоритмы кластеризации могут отнести объект данных частично к нескольким кластерам. Степень принадлежности к нечёткому кластеру зависит от близости объекта к центру кластера. Самым популярным нечётким алгоритмом кластеризации является нечёткое С-среднее (FCM). Особенностью этого алгоритма является случайный выбор центральных точек, что заставляет итеративный процесс легко попадать в локальные оптимумы. Для решения этой задачи в последнее время успешно применяются эволюционные алгоритмы, такие как генетические, имитация отжига, муравьиные колонии и роя частиц.

Оптимизация роя частиц (PSO) – это средство оптимизации, основанное на популяции, которое может быть реализовано и легко применено для решения различных задач функциональной оптимизации. В данной статье предлагается алгоритм нечёткой кластеризации, основанный на FCM и FPSO , называемый FCM-FPSO. Результаты эксперимента на двух наборах данных показывают, что алгоритм FCM-FPSO превосходит FCM и FPSO алгоритмы.

В [1], для преодоления недостатков нечёткого С-среднего, рассмотрен нечёткий алгоритм кластеризации основанный на PSO. Предложенный алгоритм использует возможности глобального поиска в алгоритме PSO, чтобы преодолеть недостатки FCM.

В [2], предложен генетический нечёткий К-режимный алгоритм для кластеризации наборов данных. Авторы отнеслись к нечёткой кластеризации, как к задаче оптимизации, и использовали генетический алгоритм для решения проблемы с целью получения глобального оптимального решения. Для ускорения процесса сходимости алгоритма они использовали одношаговый нечёткий К-режимный алгоритм в операторе перехода, вместо традиционного оператора перехода.

В [3], предложили гибридный алгоритм кластеризации данных на основе PSO и KHM , который использует достоинства обоих алгоритмов. Предлагаемый метод позволяет не только избежать попадания KHM в локальные оптимумы, но и убирает проблему медленной сходимости алгоритма PSO.

В [4], авторы используют нечёткий алгоритм С-среднего, основанный на Пикараитерации и PSO (PPSO-FCM), для преодоления недостатков FCM алгоритма.

Алгоритм нечетких С-средних

Шаг 1. Установить параметры алгоритма: c - количество кластеров; m - экспоненциальный вес; ε - параметр останова алгоритма.

Шаг 2. Случайным образом сгенерировать матрицу нечеткого разбиения $\mathbf{F}$.

Шаг 3. Рассчитать центры кластеров:

$$V_i = \frac{\sum_{k=\overline{1,N}} (\mu_{ki})^m \cdot X_k}{\sum_{k=\overline{1,N}} (\mu_{ki})^m}, \quad i = \overline{1,c}.$$

Шаг 4. Рассчитать расстояния между объектами из $\mathbf{X}$ и центрами кластеров:

$$D_{ki} = \sqrt{\|X_k - V_i\|^2}, \quad k = \overline{1,M}, \quad i = \overline{1,c}.$$

Шаг 5. Пересчитать элементы матрицы нечеткого разбиения ($i = \overline{1,c}$, $k = \overline{1,M}$):

если $D_{ki} > 0$:

$$\mu_{ki} = \frac{1}{\left(D_{ik}^2 \cdot \sum_{j=1,c} \frac{1}{D_{jk}^2} \right)^{1/(m-1)}}$$

если $D_{ki} = 0$:

$$\mu_{kj} = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}, \quad j = \overline{1,c}.$$

Шаг 6. Проверить условие $\|F - F^*\|^2 < \varepsilon$, где F^* - матрица нечеткого разбиения на предыдущей итерации алгоритма. Если "да", то перейти к шагу 7, иначе - к шагу 3.

Шаг 7. Конец.

В приведенном алгоритме самым важным параметром является количество кластеров (c). Правильно выбрать количество кластеров для реальных задач без какой-либо априорной информации о структурах в данных достаточно сложно. Существует два формальных подхода к выбору числа кластеров.

Вторым параметром алгоритма кластеризации является экспоненциальный вес (m). Чем больше m, тем конечная матрица нечеткого разбиения F становится более "размазанной", и при $m \to \infty$ она примет вид $F = [1/c]$, что является очень плохим решением, т. к. все объекты принадлежат ко всем кластерам с одной и той же степенью. Кроме того, экспоненциальный вес позволяет при формировании координат центров кластеров усилить влияние объектов с большими значениями степеней принадлежности и уменьшить влияние объектов с малыми значениями степеней принадлежности. На сегодня не существует теоретически обоснованного правила выбора значения экспоненциального веса. Обычно устанавливают $m = 2$.

Алгоритм оптимизации роя частиц

В FPSO алгоритме позиция частицы, отображается нечётким отношением набора данных к набору центров кластеров. Значения X можно выразить следующим образом:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & & & \\ x_{M1} & x_{M2} & \cdots & x_{Mn} \end{bmatrix}$$

В каждой позиции ij располагается функция принадлежности i обьекта к j кластеру с установленными ограничениями. Поэтому матрица позиций для каждой частицы такие же, как нечёткая матрица в FCM алгоритме.

Частота для каждой частицы устанавливается используя матрицу, размер которой N*C . Для обновления позиций и частот частиц используем следующие выражения, основанные на матричных операциях.

V(t+1) = w*V(t)+(c1*r1)*(pbest(t)-X(t))+(c2*r2)*(gbest(t)-X(t))

X(t+1) = X(t)+V(t+1)

В FPSO алгоритме, как и в других алгоритмах, нам нужна функция для оценки обобщенных решений, называемая функцией пригодности.

$$\sum_{i=1,c} \sum_{k=1,N} (\mu_{ki})^m \cdot \|V_i - X_k\|^2$$

Рассмотрим алгоритм, реализующий данный подход.

Шаг 1. Инициализировать параметры размера включения популяции P,C1,C2,W и максимальное число итераций.

Шаг 2. Создать популяцию с P частицами

Шаг 3. Инициализировать X,V,pbest для каждой частицы и gbest для всего роя.

Шаг 4. Вычислить центры кластеров для каждой частицы

Шаг 5. Вычислить пригодные значения для каждой частицы

Шаг 6. Вычислить pbest для каждой частицы

Шаг 7. Вычислить gbest для всей спопуляции

Шаг 8. Обновить матрицу частот для каждой частицы

Шаг 9. Обновить матрицу позиций для каждой частиц

Шаг 10. Если условие остановки не выполнено, к шагу 4.

Условие остановки в предлагаемом методе – достижение максимального числа итераций или не улучшение gbest.

Гибридный нечёткий алгоритм С-среднего и оптимизации роя частиц для кластеризации.

FCM алгоритм быстрее FPSO алгоритма, так как вычисляется меньше функций. Однако он обычно попадает в локальные минимумы. В нашем случае FCM алгоритм объединится с FPSO алгоритмом для формирования гибридного алгоритма кластеризации, называемого FCM-FPSO, который сохранит достоинства обоих алгоритмов. FCM-FPSO алгоритм применяет FCM для частиц в стае на каждой итерации так , что пригодное значение для каждой частицы улучшается.

Шаг 1. Инициализировать параметры размера включения популяции P,C1,C2,W и максимальное число итераций.

Шаг 2. Создать популяцию с P частицами

Шаг 3. Инициализировать X,V,pbest для каждой частицы и gbest для всей популяции

Шаг 4. FPSO алгоритм:

a) вычислить центры кластеров для каждой частицы

b) вычислить пригодное значение для каждой частицы

c) вычислить pbest для каждой частицы

d) вычислить gbest для всего роя

e) обновить матрицу частот для каждой частицы

f) обновить матрицу позиций для каждой частицы

g) если условие остановки не выполнено, то к шагу 4

Шаг 5. FCM алгоритм:

a) вычислить центры кластеров для каждой частицы

b) вычислить расстояние Эвклида для каждой частицы

c) пересчитать функцию принадлежности

d) вычислить gbest для роя

e) если условие остановки FCM не выполнено, то к шагу 5

Шаг 6. Если условие остановки FCM-FPSO не выполнено, то к шагу 4.

Оценка результатов. Энтропия. В качестве оценки результата кластеризации берут функцию пригодности. В наших экспериментах мы будем использовать еще один параметр оценки. Энтропия известна как численное выражение упорядоченности системы. Энтропия разбиения достигает минимума при наибольшей упорядоченности в системе, поэтому мы можем использовать этот параметр для оценки качества кластеризации. В [5] описан способ вычисления модифицированной энтропии, в котором конечный результат не зависит от количества кластеров.

Выбор данных. В качестве тестовых данных возьмем набор точек (бабочка) и набор городов России. В первом случаи это 9 объектов с двумя характеристиками, а во втором 17 городов с шестью характеристиками.

Вычислительный эксперимент

Результаты работы процедур кластеризации для точек типа 1 при разбиение на 3 кластера.

	FCM			**FPSO**			**FCM-FPSO**		
	худшее	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**
Набор (17,2,6)	121	118	113	91.9	84.4	73.7	81.2	75.9	70.2
Энтропия	0.32	0.24	0.17	0.11	0.05	0	0.07	0.03	0

Результаты работы процедур кластеризации для точек типа 1 при разбиение на 2 кластера.

	FCM			**FPSO**			**FCM-FPSO**		
	худшее	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**
Набор (17,2,6)	117	113	110	98.9	94.4	88.7	91.2	85.9	80
Энтропия	0.5	0.4	0.3	0.2	0.15	0.1	0.15	0.12	0.09

Результаты работы процедур кластеризации для точек типа 2 при разбиение на 2 кластера.

	FCM			**FPSO**			**FCM-FPSO**		
	худшее	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**
Набор (17,2,6)	1821	1717	1698	291	243	189	270	190	147
Энтропия	0.5	0.44	0.37	0.19	0.15	0.12	0.19	0.13	0.1

Результаты работы процедур кластеризации для точек типа 2 при разбиение на 3 кластера.

	FCM			**FPSO**			**FCM-FPSO**		
	худшее	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**	**худшее**	**среднее**	**лучшее**
Набор (17,3,6)	1761	1712	1639	280	221	149	261	160	117
Энтропия	0.45	0.34	0.27	0.09	0.05	0.02	0.08	0.03	0.01

Результаты работы процедур кластеризации для точек типа 2 при разбиение на 4 кластера.

	FCM			**FPSO**			**FCM-FPSO**		
	худ-шее	**сред нее**	**Луч шее**	**худ-шее**	**сред нее**	**луч шее**	**худ-шее**	**сред нее**	**луч-шее**
Набор (17,4,6)	1741	1710	1621	278	211	139	260	159	112
Энтропия	0.45	0.34	0.27	0.09	0.05	0.02	0.08	0.03	0.01

Основываясь на результатах экспериментов, можно сказать, что эти алгоритмы работают лучше при следующих настройках: c1=2, c2=2 ,P = 3 ,w = 0,9.

Заключение

Алгоритм нечёткого К-среднего чувствителен к инициализации и легко попадает в локальные оптимумы. В данной статье, для предотвращения недостатков алгоритма нечёткого К-среднего, предложено объединить его с алгоритмом нечёткого роя частиц. Эксперименты на двух наборах данных показали, что предлагаемый гибридный метод является эффективным и может показывать хорошие результаты с учётом качества найденного решения и энтропии.

**Работа выполнена при финансовой поддержке РФФИ, проекты №11-01-00359-а; 12-07-000493*

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1) Lili Li, Xiyu Liu, Mingming Xu. A novel fuzzy clustering based on particle swarm optimization. First IEEE International Symposium on Information Technologies and Applications in Education, 2007. ISITAE '07. pp. 88-90.

2) Gan G., Wu J., Yang Z. A genetic fuzzy K-modes algorithm for clustering categorical data. Expert Systems with Applications: Vol. 36 Issue 2, March, 2009, pp. 1615-1620.

3) Yang F., Sun T., Zhang C. An efficient hybrid data clustering method based on K-harmonic means, and particle swarm optimization. Expert Systems with Applications: Vol. 36 Issue 6, August, 2009 pp 9847-9852

4) Liu H.C., Yih J.M., Wu D.B., Liu S.W. Fuzzy C-means clustering algorithm based on Picard iteration and particle swarm optimization. In 2008 international workshop on education technology and training 2008 international workshop on geosciences and remote sensing. pp. 838-842.

5) Барсегян А А .Анаиз данных и процессов /А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. -3-е изд. перераб. и доп.-СПб : БХВ-Петербург, -2009.-512с.