УДК 616.12

Л.Г. Ахметшина, А.А. Егоров, Т.С. Ямнич

ВЛИЯНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ В ЗАДАЧЕ НЕЧЕТКОЙ ИНТЕРПОЛЯЦИИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Аннотация. Исследовано влияние методов кластеризации и вида функции принадлежности на результаты нечеткой интерполяции экспериментальных пространственных данных, заданных на неравномерной сетке.

Ключевые слова: пространственные данные, интерполяция, неравномерная сетка, нечеткие модели, кластеризация, функция принадлежности.

Постановка проблемы. При анализе неравномерно распределенных пространственных экспериментальных данных, практически всегда возникает проблема восстановления (интерполяции) их значений в узлах регулярной сетки. Например, моделирование рельефа и изолиний Земной поверхности, оценка границ возможных зон подтопления и загрязнения, оценка рудного тела, вычисление объемов под объектами на поверхности и т.д. по результатам полевых работ. Данные задачи имеют большую практическую значимость, и их решению уделяется большое внимание.

Наиболее известными методами, обеспечивающими восстановление полей признаков по дискретным данным в узлах регулярной сетки, можно считать методы минимальной кривизны, триангуляции, ближайшего соседа, кригинга, радиальных базисных функций, обратного расстояния в степени. Возможность использования различных нечетких моделей для решения подобных задач рассмотрены в [2], однако этот подход успешно применяется в системах контроля, но почти отсутствуют работы по применению нечеткой интерполяции для решения задач прогнозирования и классификации.

 ISSN 1562-9945

Преимущество нечетких моделей связано с возможностью использования значительно меньших объемов исходной информации, причем она может носить приближенный характер.

Существует два принципиально различных подхода при построении нечетких моделей – на основе экспертных знаний о системе и на основе измерений входов и выходов. В последнем случае различают самоорганизующиеся, адаптивные и частотные методы. Обычно, модели, построенные частотными методами (методами на основе кластеризации), используются, если функционирование системы в основном в области кластеров которым и соответствуют наблюдаемые входные значения, а возникновение других состояний носит спорадический характер [1].

В работе [2] демонстрируется возможность интерполяции экспериментальных пространственных данных, заданных на неравномерной сетке методом двумерного проецирования нечетких кластеров, в котором при определении поверхности решения используются результаты кластеризации для случая, когда исходные экспериментальные данные распределены случайным образом и имеют равнозначное значение. Основными параметрами, оказывающими непосредственное влияние на качество модели, полученной данным методом, являются алгоритм кластеризации и форма функции принадлежности.

Целью данной работы является демонстрация влияния метода кластеризации и способа формирования функции принадлежности, используемых в методе нечеткого проецирования кластеров, на точность интерполяции пространственных данных, заданных на неравномерной сетке.

Основная часть. Метод двумерного проецирования нечетких кластеров реализует идею [1], заключающуюся в предположении, что если образец принадлежит кластеру рассматриваемого класса в n-мерном пространстве, то его проекции также принадлежат проекциям этого кластера на любое из двумерных пространств $X_i \times X_j$, а функцию принадлежности заданного класса $u(x_1, x_2, ..., x_n)$ можно определить через функции принадлежности его проекций на отдельные подпространства $u(x_1, x_2), ...u(x_1, x_n), ...u(x_{n-1}x_n)$.

Алгоритм метода включает последовательность следующих действий [2].

1. Кластеризация неравномерных измерений N на c классов.

2. Построение проекций $m_i^{X_j}$ центров кластеров на входное подпространство $X_i \times X_l$, где $j,l = 1,...n-1$, $i = 1,...,c$ и определение в них функций принадлежности u_{ij}^X для каждой точки исходных данных, причем n-ый параметр – целевой, интерполяция которого производится.

3. Определение функций принадлежности полного координатного пространства интерполируемой поверхности, которое выполняется в соответствии с выражением:

$$u_i^X(x_j) = \left(\sum_{k=1}^{c} \left[\frac{d_{ij}^X(x)}{d_{kj}^X} \right]^{\frac{2}{q-1}} \right)^{-1}, \; d_{ij}^X(x) = \left| x - m_i^X \right| = \sqrt{\left(x_j - m_i^X \right)^T A \left(x_j - m_i^X \right)}, \quad (1)$$

где $A = I$ – единичная матрица. Расстояние $d_{ij}^X(x)$ между вектором x_i и центром кластера m_i^X определяется только во входном пространстве.

3. Вычисление выхода нечеткой модели (интерполируемой поверхности) y по формуле

$$y(x_j) = \left(\sum_{i=1}^{c} m_i^Y \cdot u_i^X(x_j) \right) / \left(\sum_{i=1}^{c} u_i^X(x_j) \right) \quad (2)$$

Этап кластеризации, обеспечивает разбиение множества входных значений $X = \{x_i,\ldots,x_n\} \subset \Re^p$ на $c \in \{2,\ldots,n-1\}$ подмножеств, представляющих подструктуру X.

Разбиение отражает особенность их распределения в пространстве и может быть описано при помощи матрицы U размером $c \times n$, каждый элемент которой u_{ik}, $i = 1,\ldots,c$, $k = 1,\ldots,n$ представляет принадлежность образца $x_k \in X$ к i-му кластеру. Цель алгоритма – получение векторов, определяющих значения центров классов (центроидов), характеризующих каждую группу.

Используемый метод кластеризации влияет на глобальные характеристики поверхности решения, поскольку определяет месторасположение функций принадлежности, представляющих «образ» каж-

 ISSN 1562-9945

дого из классов (центроид – центр функции принадлежности соответствующего класса), на основании которых производится расчет отображения «вход-выход» системы в нечеткой модели.

При определении множества матриц разбиения рассматриваются четкие, нечеткие и возможные (вероятностные) декомпозиции, которые трактуются как модификации нечеткой кластеризации ($M_{hcm} \subset M_{fcm} \subset M_{ncm} \subset M_{pcm} \subset [0,1]^{cn}$) [3]:

– множество с четким разбиением

$$M_{hcm} = \{ U \in \{0,1\}^{cn} \left| \sum_{i=1}^{c} u_{ik} = 1, k = 1,\ldots,n, \sum_{k=1}^{n} u_{ik} > 0, i = 1,\ldots c \right. \}$$

– множество с нечетким разбиением

$$M_{fcm} = \{ U \in [0,1]^{cn} \left| \sum_{i=1}^{c} u_{ik} = 1, k = 1,\ldots,n, \sum_{k=1}^{n} u_{ik} > 0, i = 1,\ldots c \right. \}$$

– множество с зашумленным нечетким разбиением

$$M_{ncm} = \{ U \in [0,1]^{cn} \left| 0 < \sum_{i=1}^{c} u_{ik} \le 1, k = 1,\ldots,n, \sum_{k=1}^{n} u_{ik} > 0, i = 1,\ldots c \right. \}$$

– множество с вероятностным разбиением

$$M_{pcm} = \{ U \in [0,1]^{cn} \left| \sum_{i=1}^{c} u_{ik} > 0, k = 1,\ldots,n \right. \}.$$

При интерпретации результатов нечеткой кластеризации дефаззификация обычно осуществляется на основе максимума функции принадлежности. Такой подход является не совсем корректным, поскольку значения функции принадлежности могут иметь несколько экстремумов сопоставимой или даже равной амплитуды, а значения функции принадлежности каждого из классов, содержит информацию, пригодную для анализа о степени влияния каждой точки на остальные.

В [2] при выполнении первого этапа алгоритма использовался широко известный метод нечеткой кластеризации FCM (fuzzy c-means), который базируется на процедуре итеративной минимизации целевой функции вида

$$J(U,V) = \sum_{i=1}^{c} \sum_{n=1}^{N} u_{ik}^{m} \left\| x_n - v_i \right\|^2, \qquad (3)$$

причем $\sum_{i=1}^{c} u_{ik} = 1, k = 1,\ldots,n$, где $V = \{v_1,...,v_c\}$ – центры кластеров (центроиды); $u_{ik} - i-$я функция принадлежности k-го входа x_k, $m \in [1,\infty)$ – параметр фаззификации или экспоненциальный вес, определяющий степень нечеткости результата кластеризации.

Этот метод позволяет получить хорошие результаты в случаях, когда значения выборок нескольких классов близко расположены, или даже перекрываются. К недостаткам алгоритма FCM относятся необходимость априорного задания количества кластеров, неустойчивая реакция на выбросы и артефакты, отсутствие возможности учета пространственной составляющей.

В настоящее время существуют различные модификации методов нечеткой кластеризации, направленные на повышение достоверности группирования при решении специфических задач [3]. Например, при наличии одиночных объектов (точек с низкой принадлежностью к каждому кластеру) рекомендуется применение вероятностного алгоритма c- means (PCM), который минимизирует целевую функцию

$$J_{PCM}(U,V,M) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik}\left\|x_k - v_i\right\|^2 + \eta_i(1-u_{ik})^m), \quad (4)$$

где η_i- некоторая положительная константа, характеризующая ширину возможного распределения, и не использует условие нормализации $\sum_{i=1}^{c} u_{ik} = 1, k = 1,\ldots,n$. В алгоритме Густафсона-Кесселя (GK) и Fuzzy Maximum Length Estimates (FMLE) матрица, используемая при расчете расстояния $A_i \neq I$, что позволяет находить кластера произвольных форм.

В [4] предложен алгоритм FCM-SOM, в котором сочетается применение нечеткого и нейросетевого подходов – используется самоорганизующаяся карта Кохонена (СОК) на этапе дефаззификации, что обеспечивает реорганизацию полученных кластеров и адаптивное определение их количества.

Алгоритм применения FCM-SOM в методе двумерного проецирования нечетких кластеров предлагает выполнение следующих шагов.

1. Задание числа c – количества нечетких кластеров, использование алгоритма FCM для проведения процедуры нечеткой кластеризации с целью определения матрицы U – значений функций принадлежности для каждой точки входного пространства в соответствии с (1) и центров кластеров v на основе выражения

$$v_{kj} = \frac{\sum_{i=1}^{n} \left(u_{ki}\right)^m \cdot x_{ij}}{\sum_{i=1}^{n} \left(u_{ki}\right)^m} \left(\forall k \in \{1,...,c\}, \forall j \in \{1,...,q\}\right)$$

где q – количество информативных признаков, описывающих каждый экземпляр данных.

2. Формирование данных для кластеризации с использованием СОК, которые вместо интерполируемого значения содержат значения функций принадлежности u_{ij}^{X} каждой точки исходных данных и центров кластеров.

3. Кластеризация данных с использованием СОК.

4. Определение функций принадлежности полного координатного пространства интерполируемой поверхности на основе результатов кластеризации СОК.

5. Нормализация полученных функций принадлежности полного координатного пространства.

6. Вычисление выхода нечеткой модели (интерполируемой поверхности) y по формуле (2).

Экспериментальные результаты были получены на модельных и реальных данных. В экспериментах использовался параметр фаззификации $m = 2$, пороговое значение $\varepsilon = 10^{-5}$ СОК, архитектура которой выбиралась с учетом того, что на результаты существенное влияние оказывает как общее число нейронов в сети, которое ограничивает максимально возможное значение кластеров, так и число нейронов n_x, n_y по осям x и y, соответственно.

На рис. 1 а изображен модельный набор неравномерно распределенных трехмерных данных, состоящий из 40 точек, принадлежащим двум объектам (выбор точек осуществлялся случайным образом) и расположение проекции центров кластеров, получаемых различными методами нечеткой кластеризации, на плоскость XoY. В табл. 1

приведены значения для центроидов и функции принадлежности одной из точек, расположенной в пограничной для двух объектов области.

Часто вычислительные ограничения не позволяют проанализировать все возможные декомпозиции данных, n объектов на c групп, и оценить их влияние на конечную ошибку, поэтому важным является выбор алгоритма или группы алгоритмов, наиболее подходящих для решения конкретной задачи.

Проведенные эксперименты по интерполяции пространственных данных показали, что на точность нечеткой модели в случае, когда экспериментальная выборка имеет небольшие размеры (1-5%), распределена случайным образом и расстояние между отдельными измерениями существенно различаются и также имеются внутренние области, не покрытые ими, большее влияние оказывает не метод кластеризации, а способ формирования функции принадлежности.

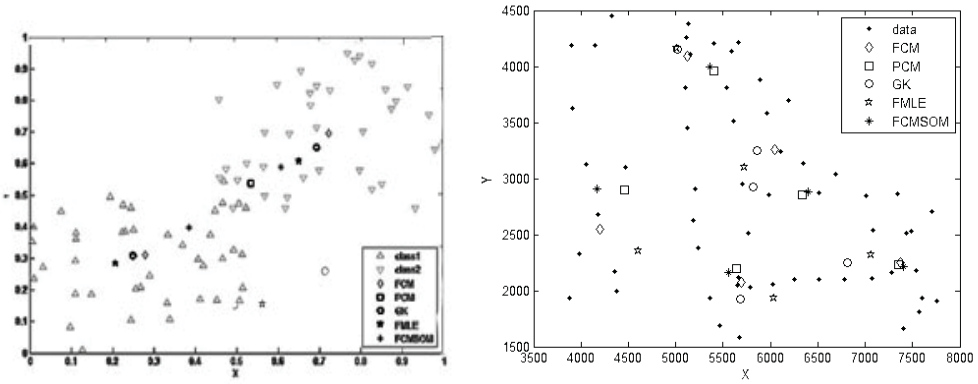


Рисунок 1 - Расположение центров кластеров при использовании различных методов кластеризации

Использование экспоненциальных функций принадлежности в соответствии с выражением (1) является источником систематической ошибки,

поскольку они имеют локальные экстремумы, находящихся на значительном расстоянии от центров кластеров, что противоречит первому закону географии.

 ISSN 1562-9945

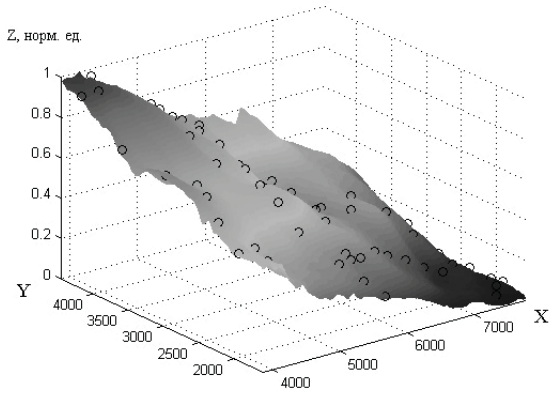
Таблица 1

Центры кластеров и функция принадлежности «проблемной» точки

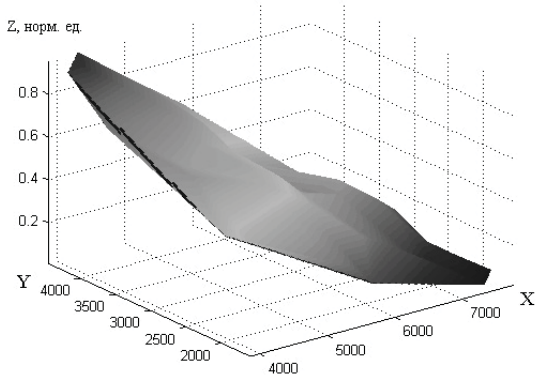
	(0,4902;0,4574;0,3470)		center1			center2		
метод	class1	class2	x	y	z	X	y	z
FCM	0,4454	0,5970	0,2784	0,3100	0,6404	0,7219	0,6957	0,4563
PCM	0,5960	0,8737	0,5333	0,5379	0,5341	0,5331	0,5379	0,5341
GK	0,8191	0,3140	0,2489		0,5478	0,6923	0,6498	0,5478
FMLE	0,2630	2,6059e-05	0,2069	0,2830	0,5238	0,6498	0,6080	0,5547
FCM-SOM	0,4197	0,6231	0,3845	0,3965	0,5949	0,6065	0,5898	0,5025

Этот недостаток устраняется при использовании метода нечеткого проецирования кластеров на основе FCM-SOM.

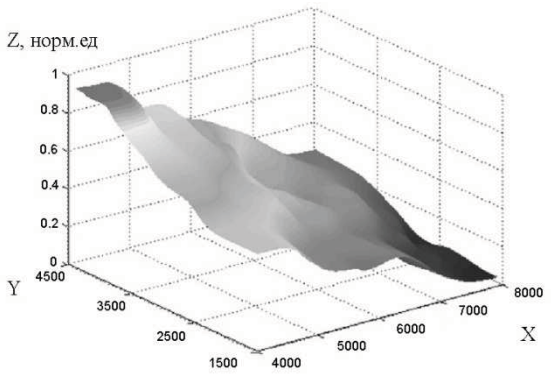
На рис. 2 а представлено изображение гравитационного поля участка поверхности Земли.



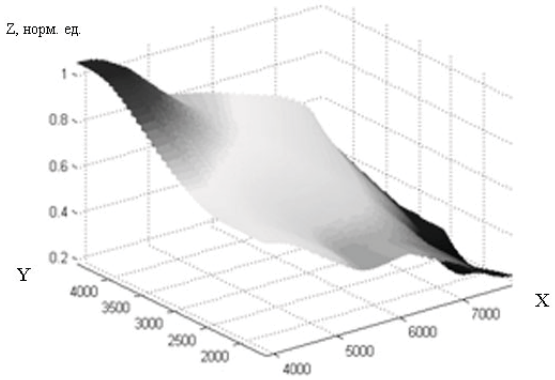
а



б



в



г

Рисунок 2 -Интерполяция геофизического поля (а); б – триангуляция Делоне; в, г – метод двумерного проецирования нечетких кластеров с использованием методовFCM и FCM-SOM, соответственно

В качестве исходных для интерполяции данных использовались значения Z, определенные для точек, изображенных на рис. 2 а, координаты (x, y) которых, соответствуют 57 реальным разведочным скважинам. Интерполяция осуществлялась при использовании различных методов кластеризации и способах построения функций принадлежности.

Выводы.

1. Точность моделирования реальных систем на основе экспериментальных данных существенно зависит от их пространственного распределения, при этом единственным способом оценки правдоподобности процедуры восстановления является эмпирическое тестирование модели.

2. При малом значении экспериментальных данных, распределение которых носит характер, близкий к случайному, и значительном увеличении расстояния между ними, влияние метода нечеткой кластеризации на глобальные характеристики поверхности решения уменьшается. Большее значение имеет выбор вида функции принадлежности, на основе которых производится расчет выхода модели.

3. Описанный подход применим для прогнозирования значений любых пространственных данных, и имеет значительный потенциал для дальнейшего развития.

ЛИТЕРАТУРА

1. А. Пегат. Нечеткое моделирование и управление. – М.: «Бином». 2009, 798 с.
2. Л.Г. Ахметшина., Т.С. Ямнич. Интерполяция пространственных данных методом двумерного проецирования нечетких кластеров // Искусственный интеллект, –2010. № 3. – С. – 433 –438.
3. Л.Рутковский. Методы и технологии искусственного интеллекта. – М.: «Горячая линия-телеком». 2010, 520 с.
4. Ахметшина Л.Г. Многопараметровый анализ изображений геофизических полей на основе комбинации алгоритмов нечеткой сегментации и нейронной сети Кохонена / Ахметшина Л.Г. // Науковий вісник Національного гірничого університету. – 2004. – № 10. – С. 44-47.

 ISSN 1562-9945