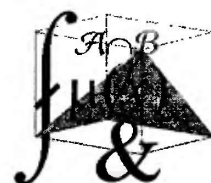


Глава 13



Нечеткая кластеризация в Fuzzy Logic Toolbox

13.1. Общая характеристика задач кластерного анализа

Термином *кластерный анализ* принято обозначать совокупность методов, подходов и процедур, разработанных для решения проблемы формирования однородных классов в произвольной проблемной области.

Необходимость анализа больших объемов объективной и субъективной информации, связанных с неформализуемыми и плохо формализуемыми задачами различной физической природы, потребовала интенсивного развития новых научных направлений, среди которых важную роль играют прикладная статистика и методы анализа данных.

Методология применения методов *прикладной статистики* основывается на предположении о вероятностной интерпретации анализируемой информации и получении в результате применения этих методов закономерностей, имеющих стохастический характер.

Методы анализа данных, составной частью которых являются методы кластерного анализа, напротив, не используют априорных предположений о вероятностной природе исходной информации и руководствуются только эвристическими соображениями о характере и особенностях исследуемой совокупности объектов. Разработанные в рамках данного направления методы и алгоритмы приобретают в последнее время новое содержание в связи с исследованиями по теории возможностей. В основе данной теории лежит нечетко-возможностная интерпретация неопределенности, что в значительной степени согласуется с исходными установками методологии анализа данных.

Кластерный анализ (или автоматическая классификация, распознавание образов без учителя, численная таксономия, кластер-анализ) занимает одно из центральных мест среди методов анализа данных и представляет собой совокупность подходов, методов и алгоритмов, предназначенных для нахождения некоторого разбиения исследуемой совокупности объектов на подмножества относительно сходных, похожих между собой объектов. При этом исходным допущением для выделения таких подмножеств, получивших специальное название *кластеров*,

которые иногда называют также таксонами или просто классами, служит лишь неформальное предположение о том, что объекты, относимые к одному кластеру, должны иметь большее сходство между собой, чем с объектами из других кластеров.

Таким образом, выявление или нахождение кластеров в множестве данных исследуемой совокупности должно удовлетворять следующим требованиям:

- ☐ каждый кластер должен представлять собой концептуально однородную категорию и содержать похожие объекты с близкими значениями свойств или признаков;
- ☐ совокупность всех кластеров должна быть исчерпывающей, т. е. охватывать все объекты исследуемой совокупности;
- ☐ кластеры должны быть взаимно исключаящими, т. е. ни один из объектов исследуемой совокупности не должен одновременно принадлежать двум различным кластерам.

Формально под задачей кластерного анализа заданного множества объектов понимается задача нахождения некоторого теоретико-множественного разбиения (группировки, покрытия) этого исходного множества объектов на непересекающиеся подмножества таким образом, чтобы элементы, относимые к одному подмножеству, отличались между собой в значительно меньшей степени, чем элементы из разных подмножеств. Обладающие подобным свойством подмножества также называются *кластерами*.

Возможность использования различных подходов к формальному определению кластеров послужила исходной причиной разработки большого многообразия методов и алгоритмов кластеризации. Методы и алгоритмы кластерного анализа как инструмент предварительного или разведочного анализа данных незаменимы при поиске закономерностей в больших наборах многомерных данных, таких как хранилища данных. При этом проблема кластер-анализа приобретает самостоятельное значение в контексте интеллектуального анализа данных (Data Mining).

13.2. Задача нечеткой кластеризации и алгоритм ее решения

Концептуальная взаимосвязь между кластерным анализом и теорией нечетких множеств основана на том обстоятельстве, что при решении задач структуризации сложных систем большинство формируемых классов объектов размыты по своей природе. Эта размытость состоит в том, что переход от принадлежности к непринадлежности элементов к данным классам скорее постепенен, чем скачкообразен. Поэтому наиболее адекватный ответ в подобного рода случаях следует искать не на вопрос: "Принадлежит ли рассматриваемый элемент тому или иному классу или нет?", а на вопрос: "В какой степени данный элемент принадлежит рассматриваемому классу?".

Требование нахождения однозначной кластеризации элементов исследуемой проблемной области является достаточно грубым и жестким, особенно при решении плохо или слабо структурируемых задач системного анализа. Методы нечеткой кластеризации ослабляют это требование. Ослабление требования осуществляется за счет введения в рассмотрение нечетких кластеров и соответствующих им функций принадлежности, принимающих значения из интервала $[0, 1]$.

Таким образом, в общем случае задачей нечеткой кластеризации является нахождение нечеткого разбиения или нечеткого покрытия множества элементов исследуемой совокупности, которые образуют структуру нечетких кластеров, присутствующих в рассматриваемых данных. Эта задача сводится к нахождению степеней принадлежности элементов универсума искомым нечетким кластерам, которые в совокупности и определяют нечеткое разбиение или нечеткое покрытие исходного множества рассматриваемых элементов. Ниже приводится формальная постановка задачи нечеткой кластеризации и описание одного из наиболее конструктивных алгоритмов ее решения.

Общая формальная постановка задачи нечеткого кластерного анализа

Пусть исследуемая совокупность данных представляет собой конечное множество элементов $A = \{a_1, a_2, \dots, a_n\}$, которое получило название *множество объектов кластеризации*. В рассмотрение также вводится конечное множество признаков или атрибутов $P = \{p_1, p_2, \dots, p_q\}$, каждый из которых количественно представляет некоторое свойство или характеристику элементов рассматриваемой проблемной области. При этом натуральное n определяет общее количество объектов данных, а натуральное q — общее количество измеримых признаков объектов.

Далее предполагается, что для каждого из объектов кластеризации некоторым образом измерены все признаки множества P в некоторой количественной шкале. Тем самым каждому из элементов $a_i \in A$ поставлен в соответствие некоторый вектор $x_i = (x_1^i, x_2^i, \dots, x_q^i)$, где x_j^i — количественное значение признака $p_j \in P$ для объекта данных $a_i \in A$. Для определенности будем полагать, что все x_j^i принимают некоторые действительные значения, т. е. $x_j^i \in \mathbb{R}$.

Вообще говоря, проблема измерения свойств или признаков у объектов данных является нетривиальной и имеет самостоятельное значение. В частности, процесс измерения свойств может быть реализован в различных шкалах, каждая из которых характеризуется допустимым преобразованием данных. В связи с этим для уточнения особенностей процесса измерений различают несколько *типов шкал измерений*. Ниже дается их краткая характеристика.

- **Шкала наименований** или номинальная шкала. Является наиболее простой из всех шкал измерений, поскольку может быть использована для установления отношения эквивалентности элементов относительно рассматриваемого признака. В данном случае в процессе измерения некоторого признака объекту ставится в соответствие некоторый символ или номер, который лишь отлича-

ет одно значение признака от другого. Допустимым преобразованием в шкалах наименований, которое не искажает результаты измерения признаков в этой шкале, является биективное или взаимно однозначное отображение между двумя множествами значений признаков. Бинарная шкала, которая состоит из двух элементов, обозначаемых произвольными символами, например: $\{0, 1\}$ или $\{+, -\}$, удобно считать частным случаем шкалы наименований. Примеры признаков, измеряемых в шкалах наименований, — пол человека, марка автомобиля, название улиц, городов и других объектов.

- *Порядковая шкала* или шкала порядка. В дополнение к различию элементов по значениям признаков позволяет установить отношение порядка элементов относительно рассматриваемого признака. В этом случае в процессе измерения признака объекту ставится в соответствие, как правило, некоторое натуральное или целое число, которое может быть интерпретировано как значение признака в баллах. Допустимым преобразованием в шкалах порядка, которое не искажает результаты измерения признаков в этой шкале, является произвольное монотонно возрастающее отображение или функция между двумя множествами значений признаков. Примеры признаков, измеряемых в порядковых шкалах, — баллы или оценки на экзаменах, баллы за выступления в некоторых видах спорта.
- *Интервальная шкала* или шкала интервалов. В дополнение к порядку элементов по значениям признаков позволяет установить равенство интервалов значений рассматриваемого признака. В этом случае в процессе измерения признака объекту ставится в соответствие, как правило, некоторое действительное число, равное значению этого признака. Допустимым преобразованием в шкалах интервалов является произвольная линейная возрастающая функция между двумя множествами значений признаков. Характерным свойством этой шкалы является отсутствие абсолютного нуля. Пример признака, измеряемого в интервальной шкале, — температура в шкалах Цельсия и Фаренгейта.
- *Шкала отношений*. В дополнение к равенству интервалов элементов по значениям признаков позволяет установить равенство отношений значений рассматриваемого признака. В этом случае в процессе измерения признака объекту ставится в соответствие также некоторое действительное число, равное значению этого признака. Допустимым преобразованием в шкалах отношений является произвольная линейная возрастающая функция, проходящая через нуль. Характерным свойством этой шкалы является наличие абсолютного нуля. Примеры признаков, измеряемых в шкале отношений, — расстояние в метрах и футах, масса в килограммах и фунтах, скорость в км/ч и узлах.

Возвращаясь к измерению признаков у объектов кластеризации, само множество признаков следует выбирать таким образом, чтобы все x'_j были измерены в шкале отношений или шкале интервалов. Именно в этом случае результаты нечеткой кластеризации имеют содержательную интерпретацию, адекватную проблеме нахождения нечетких кластеров.

Векторы значений признаков $x_i = (x_1^i, x_2^i, \dots, x_q^i)$ удобно представить в виде так называемой *матрицы данных* D размерности $(n \times q)$, каждая строка которой равна значению вектора x_i .

Задача нечеткого кластерного анализа формулируется следующим образом: на основе исходных данных D определить такое нечеткое разбиение $\mathfrak{A}(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ (4.39)–(4.40) или нечеткое покрытие $\mathfrak{Z}(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ (4.38) множества $\mathcal{A} = A$ на заданное число c нечетких кластеров $\mathcal{A}_k$ ($k \in \{2, \dots, c\}$), которое доставляет экстремум некоторой целевой функции $f(\mathfrak{A}(\mathcal{A}))$ среди всех нечетких разбиений или экстремум целевой функции $f(\mathfrak{Z}(\mathcal{A}))$ среди всех нечетких покрытий.

Сформулированная здесь задача нечеткого кластерного анализа является слишком общей. Для ее решения требуется дополнительно уточнить вид целевой функции и тип искомых нечетких кластеров (поиск нечеткого разбиения или покрытия).

Уточненная постановка задачи нечеткой кластеризации

Один из вариантов конкретизации задачи нечеткого кластерного анализа, для решения которой может быть использована специальная функция `fcm` системы MATLAB, основан на алгоритме ее решения методом нечетких c -средних.

Для уточнения вида целевой функции $f(\mathfrak{Z}(\mathcal{A}))$ в рассмотрение вводятся некоторые дополнительные понятия. Прежде всего предполагается, что искомые нечеткие кластеры представляют собой нечеткие множества $\mathcal{A}_k$, образующие нечеткое покрытие исходного множества объектов кластеризации $\mathcal{A} = A$, для которого условие (4.38) принимает следующий вид:

$$\sum_{k=1}^c \mu_{\mathcal{A}_k}(a_i) = 1 \quad (\forall a_i \in A), \quad (13.1)$$

где c — общее количество нечетких кластеров $\mathcal{A}_k$ ($k \in \{2, \dots, c\}$), которое считается предварительно заданным ($c \in \mathbb{N}$ и $c > 1$).

Примечание

Необходимость условия (13.1) обуславливается тем обстоятельством, что искомое нечеткое покрытие должно "покрывать" обычное не нечеткое множество объектов кластеризации A , которое в то же время является нечетким множеством $\mathcal{A}$, для которого значения функций принадлежности каждого из элементов равны 1. Иногда нечеткое покрытие, которое удовлетворяет условию (13.1), называют нечетким разбиением в смысле Заде.

Далее для каждого нечеткого кластера вводятся в рассмотрение так называемые *типичные представители* или *центры* v_k искомых нечетких кластеров $\mathcal{A}_k$

($k \in \{2, \dots, c\}$), которые рассчитываются для каждого из нечетких кластеров и по каждому из признаков по следующей формуле:

$$v_j^k = \frac{\sum_{i=1}^n (\mu_{A_k}(a_i))^m \cdot x_j^i}{\sum_{i=1}^n (\mu_{A_k}(a_i))^m} \quad (\forall k \in \{2, \dots, c\}, \forall p_j \in P), \quad (13.2)$$

где m — некоторый параметр, называемый *экспоненциальным весом* и равный некоторому действительному числу ($m > 1$). Каждый из центров кластеров представляет собой вектор $v_k = (v_1^k, v_2^k, \dots, v_q^k)$ в некотором q -мерном нормированном пространстве, изоморфном $\mathbb{R}^q$, т. е. $v_j^k \in \mathbb{R}^q$, если все признаки измерены в шкале отношений.

Наконец, в качестве целевой функции будем рассматривать сумму квадратов взвешенных отклонений координат объектов кластеризации от центров искоемых нечетких кластеров:

$$f(A_k, v_j^k) = \sum_{i=1}^n \sum_{k=1}^c (\mu_{A_k}(a_i))^m \sum_{j=1}^q (x_j^i - v_j^k)^2, \quad (13.3)$$

где m — экспоненциальный вес нечеткой кластеризации ($m \in \mathbb{R}$, $m > 1$), значение которого задается в зависимости от количества элементов (мощности) множества A . Чем больше элементов содержит множество A , тем меньшее значение выбирается для m .

Задача нечеткой кластеризации. *Задача нечеткой кластеризации* может быть сформулирована следующим образом: для заданных матрицы данных D , количества нечетких кластеров c ($c \in \mathbb{N}$ и $c > 1$), параметра m определить матрицу U значений функций принадлежности объектов кластеризации $a_i \in A$ нечетким кластерам $\mathcal{A}_k$ ($k \in \{2, \dots, c\}$), которые доставляют *минимум* целевой функции (13.3) и удовлетворяют ограничениям (13.1), (13.2), а также дополнительным ограничениям (13.4) и (13.5):

$$\sum_{i=1}^n \mu_{A_k}(a_i) > 0 \quad (\forall k \in \{2, \dots, c\}); \quad (13.4)$$

$$\mu_{A_k}(a_i) \geq 0 \quad (\forall k \in \{2, \dots, c\}, \forall a_i \in A). \quad (13.5)$$

Условие (13.4) исключает появление пустых нечетких кластеров в искомой нечеткой кластеризации. Последнее условие (13.5) имеет чисто формальный характер, поскольку непосредственно следует из определения функции принадлежности нечетких множеств. В этом случае минимизация целевой функции (13.3) минимизирует отклонение всех объектов кластеризации от центров нечетких кластеров пропорционально значениям функций принадлежности этих объектов соответствующим нечетким кластерам.

Поскольку целевая функция (13.2) не является выпуклой, а ограничения (13.1), (13.2), (13.4), (13.5) в своей совокупности формируют невыпуклое множество допустимых альтернатив, то в общем случае задача нечеткой кластеризации относится к многоэкстремальным задачам нелинейного программирования.

Достоинством постановки задачи нечеткой кластеризации в виде (13.1)—(13.5) является естественная интерпретация как искомым нечетких кластеров, определяемых функциями принадлежности (13.5), так и их типичных представителей или центров (13.2), которые также определяются в результате решения поставленной задачи.

Недостатком данной постановки задачи нечеткой кластеризации является необходимость априорного задания общего числа нечетких кластеров c ($c \in \mathbb{N}$ и $c > 1$), которое в отдельных случаях может быть неизвестно. Это обстоятельство может потребовать привлечения дополнительных процедур для его определения либо решения поставленной задачи для нескольких значений c с последующим выбором наиболее адекватного результата нечеткой кластеризации.

Алгоритм решения задачи нечеткой кластеризации методом нечетких c -средних

Основные идеи алгоритма для решения сформулированной задачи нечеткой кластеризации были предложены Дж. К. Данном (J. C. Dunn) в 1974 г. Этот алгоритм первоначально получил название нечеткого алгоритма ISODATA (fuzzy ISODATA или F-ISODATA). В 1980 г. Дж. К. Беждек (J. C. Bezdek) теоретически доказал сходимость этого алгоритма. Позже в 1981 г. Дж. Беждек обобщил алгоритм FISODATA на случай произвольных нечетких многообразий и предложил для этого алгоритма название нечетких c -средних (FCM, Fuzzy C-Means). Именно под таким названием алгоритм сейчас наиболее известен и реализован в системе MATLAB.

Алгоритм FCM для решения задачи нечеткой кластеризации в виде (13.1)—(13.5) имеет итеративный характер последовательного улучшения некоторого исходного нечеткого разбиения $\mathcal{R}(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$, которое задается пользователем или формируется автоматически по некоторому эвристическому правилу. На каждой из итераций рекуррентно пересчитываются значения функций принадлежности нечетких кластеров и их типичные представители.

Алгоритм FCM закончит работу в случае, когда произойдет выполнение заданного априори некоторого конечного числа итераций, либо когда минимальная абсолютная разность между значениями функций принадлежности на двух последовательных итерациях не станет меньше некоторого априори заданного значения.

Формально алгоритм FCM определяется в форме итеративного выполнения следующей последовательности шагов:

1. Предварительно необходимо задать следующие значения: количество искомым нечетких кластеров c ($c \in \mathbb{N}$ и $c > 1$), максимальное количество итераций алгорит-

ма s ($s \in \mathbb{N}$), параметр сходимости алгоритма ε ($\varepsilon \in \mathbb{R}_+$), а также экспоненциальный вес расчета целевой функции и центров кластеров m (как правило, $m=2$). В качестве *текущего* нечеткого разбиения на первой итерации алгоритма для матрицы данных D задать некоторое исходное нечеткое разбиение $\mathfrak{R}(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ на c непустых нечетких кластеров, которые описываются совокупностью функций принадлежности $\mu_k(a_i)$ ($\forall k \in \{2, \dots, c\}, \forall a_i \in A$).

2. Для исходного текущего нечеткого разбиения $\mathfrak{R}(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ по формуле (13.2) рассчитать центры нечетких кластеров v_j^k ($\forall k \in \{2, \dots, c\}, \forall p_j \in P$) и значение целевой функции $f(\mathcal{A}_k, v_j^k)$ по формуле (13.3). Количество выполненных итераций положить равным 1.
3. Сформировать *новое* нечеткое разбиение $\mathfrak{R}'(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ исходного множества объектов кластеризации A на c непустых нечетких кластеров, характеризующих совокупностью функций принадлежности $\mu'_k(a_i)$ ($\forall k \in \{2, \dots, c\}, \forall a_i \in A$), которые определяются по формуле:

$$\mu'_{A_k}(a_i) = \left(\sum_{l=1}^c \left(\frac{\sum_{j=1}^q (x_j^l - v_j^k)^2}{\sum_{j=1}^q (x_j^l - v_j^l)^2} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (\forall k \in \{2, \dots, c\}, \forall a_i \in A). \quad (13.6)$$

4. При этом если для некоторого $k \in \{2, \dots, c\}$ и некоторого $a_i \in A$ значение $\sum_{j=1}^q (x_j^l - v_j^k)^2 = 0$, то для соответствующего нечеткого кластера $\mathcal{A}_k$ полагаем $\mu'_k(a_i) = 1$, а для остальных $\mathcal{A}_l$ ($\forall l \in \{2, \dots, c\}, l \neq k$) полагаем $\mu'_l(a_i) = 0$. Если же таких $k \in \{2, \dots, c\}$ для некоторого $a_i \in A$ окажется несколько, т. е. для них значение $\sum_{j=1}^q (x_j^l - v_j^k)^2 = 0$, то эвристически для меньшего из k полагаем $\mu'_k(a_i) = 1$, а для остальных $l \in \{2, \dots, c\}, l \neq k$ полагаем $\mu'_l(a_i) = 0$.
5. Для нового нечеткого разбиения $\mathfrak{R}'(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ по формуле (13.2) рассчитать центры нечетких кластеров v_j^k ($\forall k \in \{2, \dots, c\}, \forall p_j \in P$) и значение целевой функции $f(\mathcal{A}_k, v_j^k)$ по формуле (13.3).
6. Если количество выполненных итераций превышает заданное число s или же модуль разности $|f(\mathcal{A}_k, v_j^k) - f(\mathcal{A}_k, v_j^l)| \leq \varepsilon$, т. е. не превышает значения параметра сходимости алгоритма ε , то в качестве искомого результата нечеткой кластеризации принять нечеткое разбиение $\mathfrak{R}'(\mathcal{A}) = \{\mathcal{A}_k \mid \mathcal{A}_k \subseteq \mathcal{A}\}$ и закончить выполнение алгоритма. В противном случае считать текущим нечетким разбиением $\mathfrak{R}(\mathcal{A}) = \mathfrak{R}'(\mathcal{A})$ и перейти на шаг 3 алгоритма, увеличив на 1 количество выполненных итераций.

Алгоритм FCM по своему характеру относится к приближенным алгоритмам поиска экстремума для целевой функции (13.3) при наличии ограничений: (13.1), (13.2), (13.4), (13.5). Поэтому в результате выполнения данного алгоритма определяется, вообще говоря, локально-оптимальное нечеткое разбиение $\mathfrak{A}^*(\mathcal{A})$, которое описывается совокупностью функций принадлежности $\mu_k(a_i)$ ($\forall k \in \{2, \dots, c\}$, $\forall a_i \in A$), а также центры или типичные представители каждого из нечетких кластеров v_j^k ($\forall k \in \{2, \dots, c\}$, $\forall p_j \in P$).

Примечание

Иногда возникает необходимость представить результаты нечеткой кластеризации в форме обычного не нечеткого разбиения. С этой целью можно использовать один из рассмотренных ранее методов дефаззификации (см. главу 7) для конечного множества значений функции принадлежности или любой другой метод приведения к четкости. Поскольку все функции принадлежности искомого нечеткого разбиения $\mu_k(a_i)$ ($\forall k \in \{2, \dots, c\}$, $\forall a_i \in A$) нормированы условием (13.1), в качестве метода дефаззификации можно использовать метод максимального значения функции принадлежности.

Опыт решения прикладных задач нечеткой кластеризации показывает, что наиболее эффективный путь получения адекватных результатов заключается в многократном выполнении алгоритма FCM для различных исходных нечетких разбиений и, если не известно количество нечетких кластеров, для различных значений c ($c \in \mathbb{N}$ и $c > 1$). Полученные результаты для одинаковых значений c сравниваются по значениям целевой функции полученных нечетких разбиений с целью принятия окончательного решения об искомой нечеткой кластеризации.

13.3. Средства решения задачи нечеткой кластеризации в пакете Fuzzy Logic Toolbox

В системе MATLAB для решения задачи нечеткой кластеризации на основе алгоритма FCM может быть использована специальная функция командной строки `fcm` или специальный графический интерфейс кластеризации, вызываемый функцией `findcluster`. Рассмотрим более подробно особенности применения этих средств.

Решение задачи нечеткой кластеризации в командном режиме

Функция командной строки `fcm` предназначена для решения задачи нечеткой кластеризации с использованием алгоритма FCM. Она может быть вызвана в одном из следующих форматов:

```
[center, U, obj_fcn] = fcm(data, cluster_n)
```

или

```
[center, U, obj_fcn] = fcm(data, cluster_n, options)
```

Входными аргументами этой функции являются:

- **data**: матрица исходных данных **D** кластеризации, i -строка которой представляет информацию об одном объекте нечеткой кластеризации $a_i \in A$ в форме вектора $x_i = (x_1^i, x_2^i, \dots, x_q^i)$, где x_j^i — количественное значение признака $p_j \in P$ для объекта данных $a_i \in A$;
- **cluster_n**: число искоемых нечетких кластеров c ($c \in \mathbb{N}$ и $c > 1$).

Выходными аргументами этой функции являются:

- **center**: матрица центров искоемых нечетких кластеров v_j^k ($\forall k \in \{2, \dots, c\}$, $\forall p_j \in P$), каждая строка которой представляет координаты центра одного из нечетких кластеров в форме вектора v_k ($\forall k \in \{2, \dots, c\}$);
- **u**: матрица значений функций принадлежности искомого нечеткого разбиения $\mu_k(a_i)$ ($\forall k \in \{2, \dots, c\}$, $\forall a_i \in A$);
- **obj_fcn**: значения целевой функции $J(\mathcal{A}_k, v_j^k)$ на каждой из итераций работы алгоритма.

Функция `fcm(data, cluster_n, options)` может быть вызвана с дополнительными аргументами `options`, которые предназначены для управления процессом нечеткой кластеризации, а также для изменения критерия остановки работы алгоритма и/или отображения информации на экране монитора.

Эти дополнительные аргументы имеют следующие значения:

- **options(1)**: экспоненциальный вес m для расчета матрицы нечеткого разбиения u (по умолчанию это значение равно $m=2$);
- **options(2)**: максимальное число итераций s (по умолчанию это значение равно $s=100$);
- **options(3)**: параметр сходимости алгоритма ϵ (по умолчанию это значение равно $\epsilon=0.00001$);
- **options(4)**: информация о текущей итерации, отображаемая на экране монитора (по умолчанию это значение равно 1).

Если любое из значений дополнительных аргументов равно `NaN` (не число), то для этого аргумента используется значение по умолчанию. Функция `fcm` заканчивает свою работу, когда алгоритм FCM выполнит максимальное количество итераций s , или когда разность между значениями целевых функций на двух последовательных итерациях будет меньше заданного априори значения параметра сходимости алгоритма ϵ .

Примечание

Функция `fcm` реализована в виде `m`-файла и использует, в свою очередь, три другие функции: функцию `initfcm` для формирования матрицы исходного разбиения некоторым случайным образом, функцию `distfcm` для расчета матрицы расстояний между точками данных и центрами кластеров и функцию `stepfcm` для значений

целевой функции и функций принадлежности объектов нечетким кластерам на каждой итерации работы алгоритма FCM. Все эти функции также реализованы в виде m-файлов и находятся в папке C:\MATLAB6p1\toolbox\fuzzy\fuzzy.

Пример 13.1. В качестве примера рассмотрим множество данных, содержащихся в системе MATLAB и использующихся в качестве тестовой совокупности объектов нечеткой кластеризации. Эти данные представляют собой матрицу данных **D** размерности (140×2), которые содержатся в файле `fcmdata.dat`, поставляемом вместе с системой MATLAB. В этом случае матрица данных **D** соответствует 140 объектам, для каждого из которых выполнены измерения по двум признакам, что является удобным для визуализации исходных данных и результатов нечеткой кластеризации в двумерном пространстве на плоскости. График координат точек на плоскости, соответствующих объектам нечеткой кластеризации, изображен на рис. 13.1.

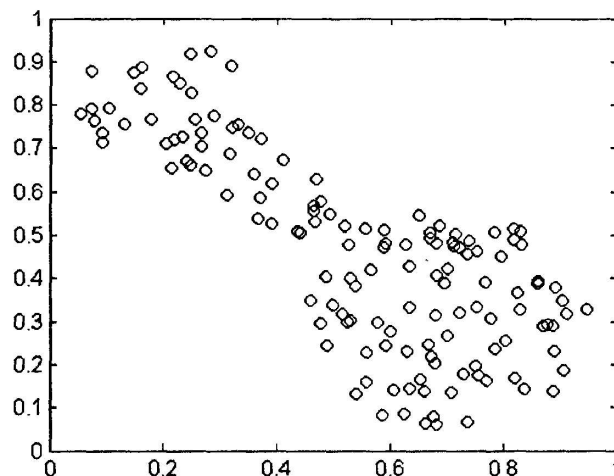


Рис. 13.1. Множество объектов нечеткой кластеризации, соответствующих матрице данных из файла `fcmdata.dat`

В листинге 13.1 приводится последовательность команд, которые обеспечивают решение задачи нечеткой кластеризации множества данных **D** и визуализацию полученных результатов. В этом примере используется первый формат записи функции `fcm`.

Листинг 13.1. Пример решения задачи нечеткой кластеризации с помощью функции `fcm` с использованием первого формата ее записи

```
load fcmdata.dat
plot(fcmdata(:,1), fcmdata(:,2), 'o', 'color', 'k')
[center, U, obj_fcn] = fcm(fcmdata, 2);
```

```

maxU = max(U);
index1 = find(U(1,:) == maxU);
index2 = find(U(2,:) == maxU);
line(fcmdata(index1,1), fcmdata(index1, 2), 'linestyle', 'none',...
'marker', 'o','color', 'g');
line(fcmdata(index2,1), fcmdata(index2, 2), 'linestyle', 'none',...
'marker', 'x','color', 'r');
hold on
plot(center(1,1),center(1,2),'ko','markersize', 10, 'LineWidth', 2)
plot(center(2,1),center(2,2),'ko','markersize', 10, 'LineWidth', 2)

```

Результат решения задачи нечеткой кластеризации для 2-х нечетких кластеров с использованием указанной последовательности команд может быть визуализирован (рис. 13.2).

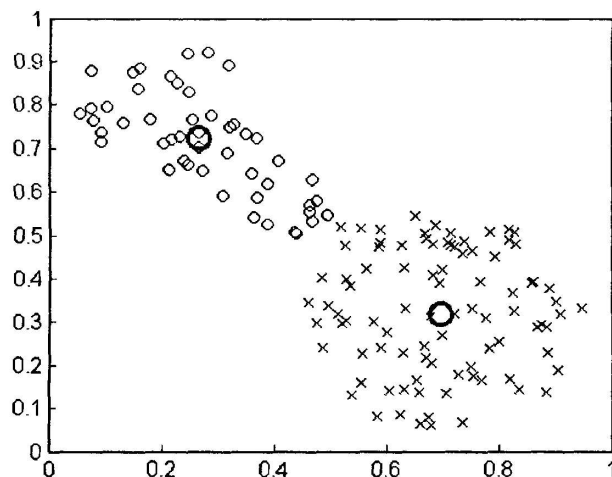


Рис. 13.2. Результат решения задачи нечеткой кластеризации для матрицы данных из файла fcmdata.dat

Если после записи функции `fcm` в третьей строке не указывать точку с запятой (;), то в окне команд будут отображены значения координат центров нечетких кластеров, значения функций принадлежности объектов нечетким кластерам и значения целевой функции на каждой из итераций работы алгоритма FCM (рис. 13.3).

Теперь выполним небольшой эксперимент и изменим параметры функции `fcm`, заданные по умолчанию. Для этого будем использовать второй формат ее записи с дополнительными значениями аргументов (листинг 13.2).

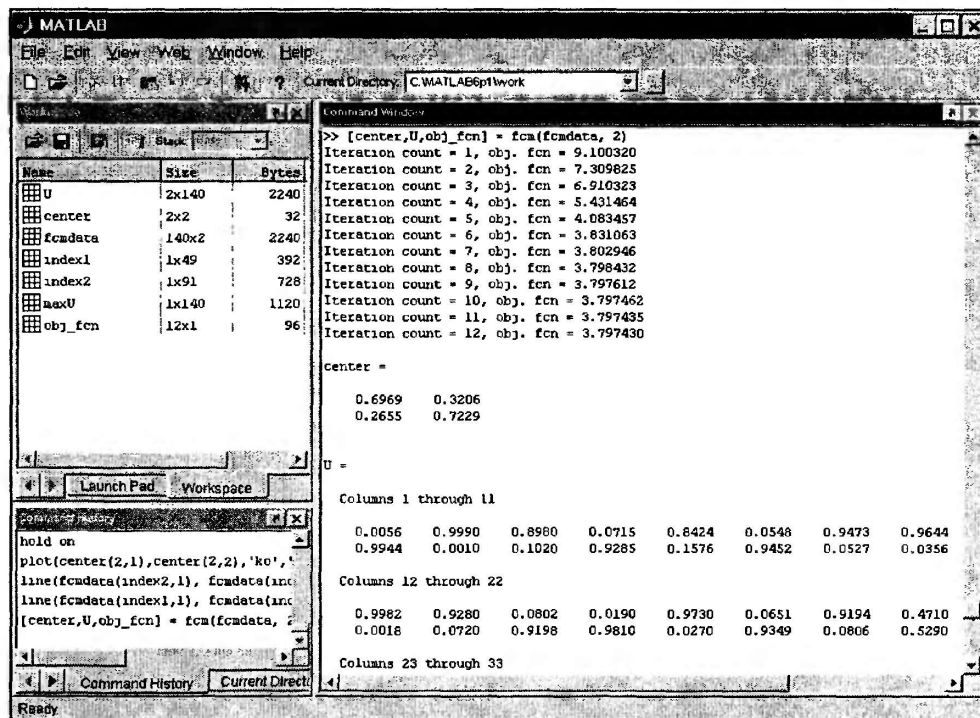


Рис. 13.3. Результаты решения задачи нечеткой кластеризации в окне команд системы MATLAB

Листинг 13.2. Пример решения задачи нечеткой кластеризации с помощью функции `fcm` с использованием второго формата ее записи

```
load fcmdata.dat
center,U,obj_fcn] = fcm(fcmdata, 2, [2.5 1000 0.0000001 1]);
maxU = max(U);
index1 = find(U(1,:) == maxU);
index2 = find(U(2,:) == maxU);
line(fcmdata(index1,1), fcmdata(index1, 2), 'linestyle', 'none',...
'marker', 'o','color', 'g');
line(fcmdata(index2,1), fcmdata(index2, 2), 'linestyle', 'none',...
'marker', 'x','color', 'r');
hold on
plot(center(1,1),center(1,2),'ko','markersize', 10, 'LineWidth', 2)
plot(center(2,1),center(2,2),'ko','markersize', 10, 'LineWidth', 2)
```

Результат решения задачи нечеткой кластеризации для 2-х нечетких кластеров с использованием указанной последовательности команд визуализирован (рис. 13.4). В этом случае экспоненциальный вес $m=2.5$, максимальное количество итераций $s=1000$, параметр сходимости алгоритма $\epsilon=0.0000001$. Как можно увидеть, столь высокая точность расчетов практически не нужна, поскольку алгоритм заканчивает свою работу после выполнения 16 итераций.

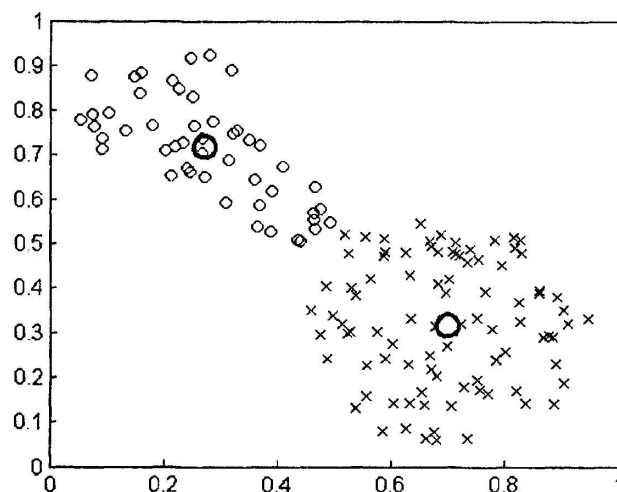


Рис. 13.4. Результат решения задачи нечеткой кластеризации с дополнительными параметрами функции `fcm`

Анализ графиков, изображенных рис. 13.2 и рис. 13.4, показывает их практическую идентичность, что позволяет сделать вывод о согласованности полученных результатов нечеткой кластеризации.

Решение задачи нечеткой кластеризации с использованием средств графического интерфейса

Функция `findcluster` предназначена для вызова графического интерфейса программы нечеткой кластеризации для метода нечетких s -средних и метода субтрактивной кластеризации (subtractive clustering). Она может быть вызвана в одном из следующих форматов: `findcluster` или `findcluster('file.dat')`.

В первом случае функция `findcluster` вызывает графический интерфейс GUI программы для выполнения нечеткой кластеризации алгоритмом FCM и/или нечеткой субтрактивной кластеризации. При этом необходимо загрузить в рабочую область исходные данные из внешнего файла с помощью кнопки **Load Data** (рис. 13.5).

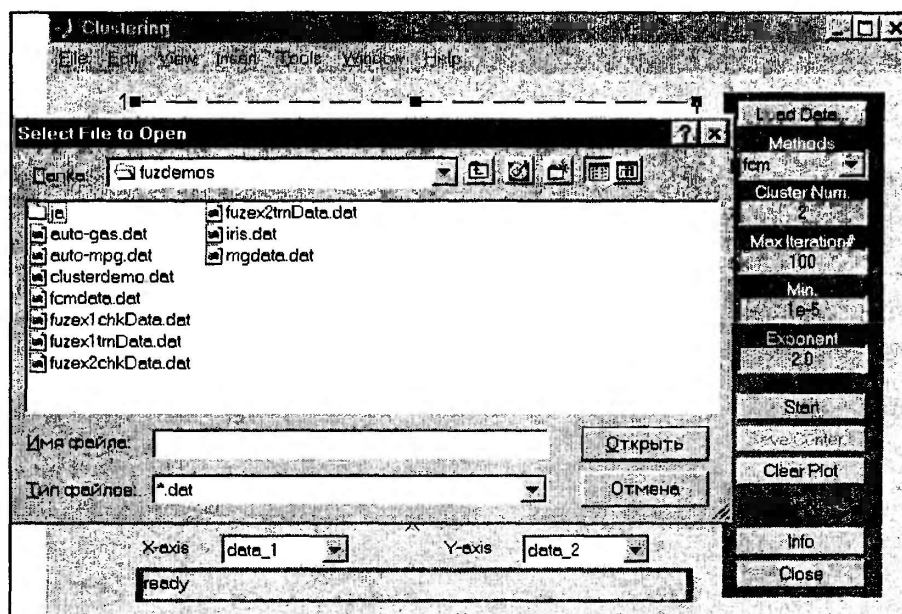


Рис. 13.5. Выбор внешнего файла для загрузки исходных данных в рабочую область

Выбор метода нечеткой кластеризации осуществляется с помощью раскрывающегося списка **Methods**. Для каждого из методов нечеткой кластеризации в соответствующих строках ввода установлены значения параметров алгоритмов по умолчанию. Эти значения могут быть изменены пользователем. Для этого необходимо установить курсор ввода в соответствующее поле и набрать нужные цифры с учетом допустимых значений параметров. Назначение этих параметров для алгоритма FCM описано ранее при рассмотрении функции *fcm*. Назначение этих параметров для алгоритма субтрактивной кластеризации будет описано ниже при рассмотрении функции *subclust*.

Во втором случае функция *findcluster('file.dat')* вызывает графический интерфейс, а в рабочую область автоматически загружаются данные кластеризации из внешнего файла *file.dat*. При этом на графике отображаются значения матрицы данных для первых двух признаков.

После нажатия на кнопку **Start** начинает работу соответствующий алгоритм нечеткой кластеризации с параметрами, установленными по умолчанию или измененными пользователем. Результаты работы алгоритма отображаются на графике (рис. 13.7). Найденные центры кластеров изображены черными кругами и их координаты можно сохранить во внешнем файле с целью последующего анализа.

Иногда число нечетких кластеров, необходимых для работы алгоритма FCM, априори является неизвестным. В этом случае целесообразно использовать реализованный в системе MATLAB так называемый алгоритм субтрактивной кластеризации.

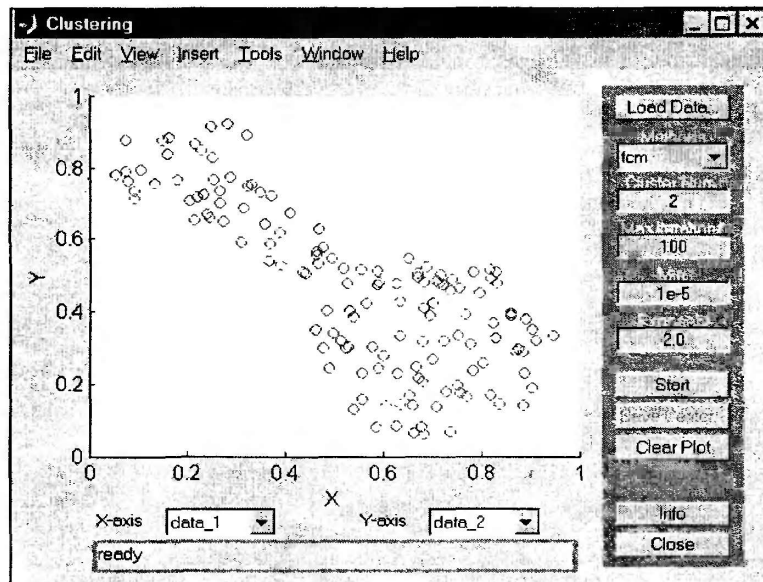


Рис. 13.6. Результат вызова графического интерфейса нечеткой кластеризации с уже загруженными исходными данными из внешнего файла fcmdata.dat

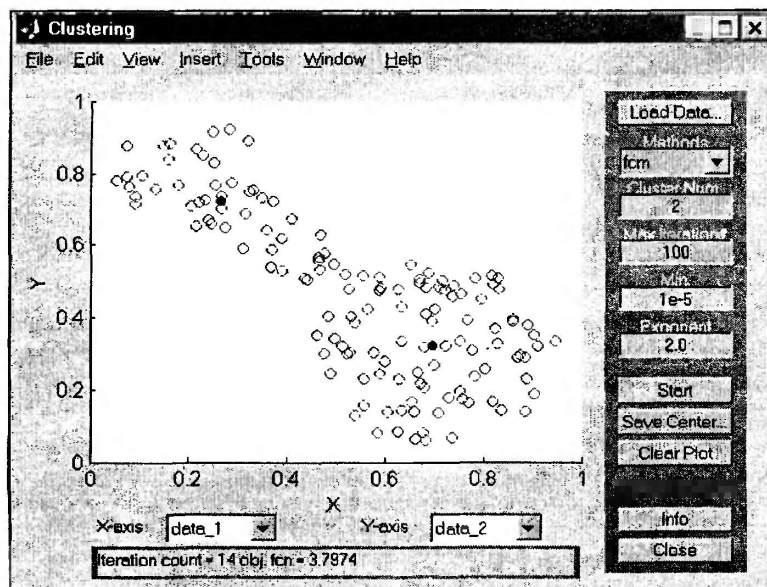


Рис. 13.7. Результат решения задачи нечеткой кластеризации алгоритмом FCM с помощью графического интерфейса для множества данных из внешнего файла fcmdata.dat

Решение задачи определения числа кластеров для нечеткой кластеризации в системе MATLAB

В случае отсутствия каких-либо априорных предположений относительно количества нечетких кластеров в системе MATLAB можно использовать функцию командной строки `subclust` или метод субтрактивной нечеткой кластеризации, реализованный в графическом интерфейсе кластеризации.

Идея метода субтрактивной кластеризации состоит в том, что каждая точка данных предполагается в качестве центра потенциального кластера, после чего следует вычислить некоторую меру способности каждой точки данных представлять центр кластера. Эта количественная мера основана на оценке плотности точек данных вокруг соответствующего центра кластера.

Данный алгоритм, который является обобщением метода кластеризации Р. Ягера (R. Yager), основан на выполнении следующих действий:

1. Выбрать точку данных с максимальным потенциалом для представления центра первого кластера.
2. Удалить все точки данных в окрестности центра первого кластера, величина которой задается параметром `radii`, чтобы определить следующий нечеткий кластер и координаты его центра.

Эти две процедуры повторяются до тех пор, пока все точки данных не окажутся внутри окрестностей радиуса `radii` искоемых центров кластеров.

Функция командной строки `subclust` находит центры кластеров методом субтрактивной кластеризации. Она используется в следующем формате:

```
[C,S] = subclust(X, radii, xBounds, options)
```

При этом матрица `X` содержит данные кластеризации, каждая строка которой соответствует координатам отдельной точки данных. Параметр `radii` представляет собой вектор, компоненты которого принимают значения из интервала $[0, 1]$ и задают диапазон расчета центров кластеров по каждому из признаков измерений. При этом делается предположение, что все данные содержатся в некотором единичном гиперкубе. В общем случае малые значения параметра `radii` приводят к нахождению малого числа больших по количеству точек кластеров. Наилучшие результаты получаются при значениях `radii` между 0.2 и 0.5.

Аргумент `xBounds` представляет собой матрицу размерности $(2 \times q)$, которая определяет способ отображения матрицы данных `X` в некотором единичном гиперкубе. Здесь q — количество рассматриваемых признаков в множестве данных. Этот аргумент является необязательным, если матрица `x` уже нормализована. Первая строка этой матрицы содержит минимальные значения интервала измерения каждого из признаков, а вторая строка — максимальные значения измерения каждого из признаков.

Для изменения заданных по умолчанию параметров алгоритма кластеризации может быть использован дополнительный вектор `options`. Компоненты этого вектора могут принимать следующие значения:

- `options(1) = quashFactor` — параметр, используемый в качестве коэффициента для умножения значений `radii`, которые определяют окрестность центра кластера. Это осуществляется с целью уменьшения влияния потенциала граничных точек, рассматриваемых как часть нечеткого кластера (по умолчанию это значение равно 1.25);
- `options(2) = acceptRatio` — параметр, устанавливающий потенциал как часть потенциала центра первого кластера, выше которого другая точка данных может рассматриваться в качестве центра другого кластера (по умолчанию это значение равно 0.5);
- `options(3) = rejectRatio` — параметр, устанавливающий потенциал как часть потенциала центра первого кластера, ниже которого другая точка данных не может рассматриваться в качестве центра другого кластера (по умолчанию это значение равно 0.15);
- `options(4) = verbose` — если значение этого параметра не равно нулю, то на экран монитора выводится информация о выполнении процесса кластеризации (по умолчанию это значение равно 0).

Функция `subclust` возвращает матрицу `C` значений координат центров нечетких кластеров. При этом каждая строка этой матрицы содержит координаты одного центра кластера. Эта функция также возвращает вектор `S`, компоненты которого представляют σ -значения, которые определяют диапазон влияния центра кластера по каждому из рассматриваемых признаков. При этом все центры кластеров обладают одинаковым множеством σ -значений.

Пример 13.2. Рассмотрим решение задачи определения количества кластеров для множества исходных данных из примера 13.1. Напомним, что эти данные представляют собой матрицу данных `D` размерности (140×2), которые содержатся в файле `fcmdata.dat`. В этом случае вызов функции субтрактивной кластеризации может быть реализован следующим образом:

```
load fcmdata.dat
[C,S] = subclust(fcmdata,[0.5 0.5],[],[1.25 0.5 0.15 1])
```

Поскольку для множества данных из примера 13.1 используется 2 признака (матрица `datain` имеет 5 столбцов), то по каждому из признаков задаются радиусы окрестностей 0.5 и 0.5 соответственно. Коэффициенты шкалирования для отображения множества данных в единичный гиперкуб получаются на основе минимального и максимального значений данных.

Аргумент `squashFactor = 1.25` указывает на то, что необходимо определить кластеры, недалеко расположенные друг от друга. Аргумент `acceptRatio = 0.5` указывает на то, что для нахождения центров кластеров очень высокий потенциал не является необходимым. Аргумент `rejectRatio = 0.15` не исключает из рассмотрения точки данных, не обладающих высоким потенциалом. Наконец,

последний аргумент `verbose = 1` позволяет вывод информации о выполнении процесса кластеризации на экран монитора.

В результате выполнения этого фрагмента команд будут получены следующие значения матрицы центров кластеров и вектора σ -значений (рис. 13.8). Как можно заметить, для указанных значений аргументов рассматриваемая функция субтрактивной кластеризации находит три нечетких кластера и отображает координаты их центров в командном окне системы MATLAB.

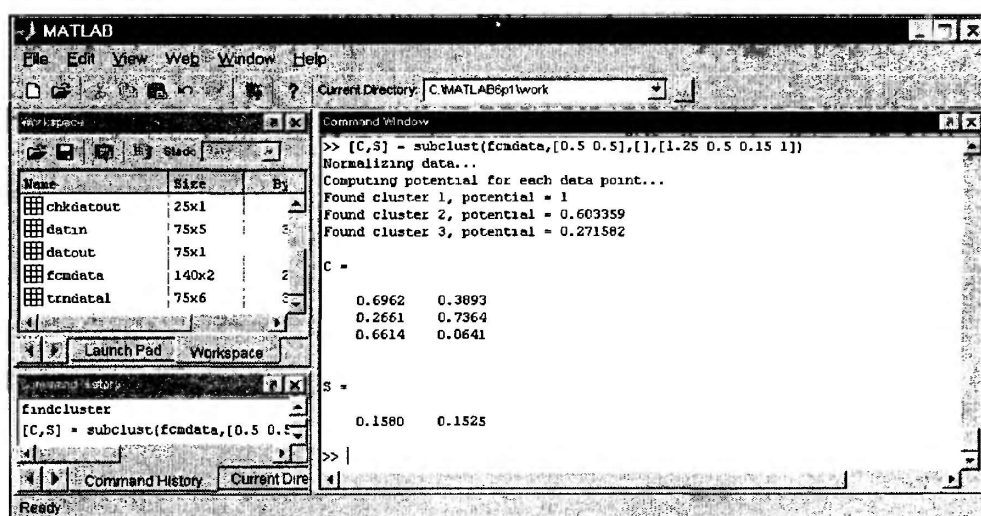


Рис. 13.8. Результат решения задачи нечеткой субтрактивной кластеризации с помощью функции командной строки для множества данных из внешнего файла `fcmdata.dat`

Для решения задачи субтрактивной кластеризации может быть также использован рассмотренный выше графический интерфейс нечеткой кластеризации, который вызывается функцией `findcluster`. Вызов этой функции может быть осуществлен в одном из следующих форматов: `findcluster` или `findcluster('file.dat')`, особенности которых были изложены ранее и иллюстрированы на рис. 13.5—13.7.

Пример 13.3. Рассмотрим решение задачи определения количества кластеров для множества исходных данных из примера 13.1 с использованием графического интерфейса кластеризации. Для этого загрузим исходные данные из внешнего файла командой `findcluster('fcmdata.dat')`, выберем метод кластеризации **subtractive** в раскрывающемся списке **Methods** и нажмем кнопку **Start**. Остальные параметры оставим предложенными по умолчанию. Результат решения задачи субтрактивной кластеризации изображен на рис. 13.9 и также содержит 3 нечетких кластера.

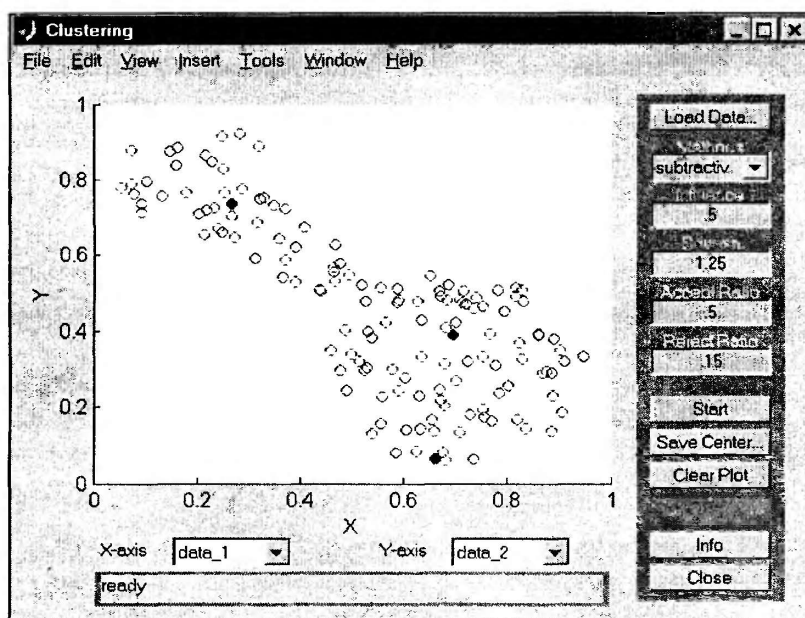


Рис. 13.9. Результат решения задачи нечеткой субтрактивной кластеризации с помощью графического интерфейса кластеризации для множества данных из внешнего файла `fcmdata.dat`

Таким образом, система MATLAB позволяет решать задачи нечеткой кластеризации двумя способами: с помощью функций командной строки и с помощью графического интерфейса кластеризации. Первый из них является более трудоемким, но обладает большей гибкостью и возможностью отображения в окне команд значений матриц центров нечетких кластеров, функций принадлежности и целевой функции. Второй способ представляется наиболее удобным для выполнения некоторой серии расчетов для различных значений входных параметров с целью визуального анализа полученных результатов.

В заключение этой главы следует отметить, что результаты нечеткой кластеризации имеют приближенный характер и могут служить лишь для предварительной структуризации информации, содержащейся в множестве исходных данных. Решая задачи нечеткой кластеризации, нужно помнить об особенностях и ограничениях процесса измерения признаков у совокупности объектов кластеризации. Поскольку нечеткие кластеры формируются на основе евклидовой метрики, соответствующее пространство признаков должно удовлетворять аксиомам метрического пространства. В то же время для поиска закономерностей в проблемной области, имеющих неметрический характер, необходимо использовать специальные средства и инструментарий, разработанные для интеллектуального анализа данных (Data Mining).