**4.1 Основные понятия и постановка задачи**

1.

Основным условием решения задач распознавания с обучением является наличие априорной информации о предметной области, в виде *обучающей выборки.*

*Кластеризация* ‒ это классификация объектов без обучения, когда процесс генерации классов происходит без априорной информации о прототипах классификации.

2.

***Формальная постановка задачи кластеризации***.

Пусть в пространстве признаков задано *m* объектов .



Кластеризацию формально можно записать следующим образом.

Определить области такие, что любой входит только в одну из областей , т.е.



для любых .



3.

Алгоритмы кластеризации включают объекты в классы (кластеры) в соответствии с *мерой сходства*, которая отражает *естественные связи* между объектами.

Предполагается, что степень этих связей намного выше среди объектов внутри группы и ниже среди объектов из разных групп.

Поэтому для классификации на основе кластерного анализа необходимо ввести меру сходства или различия.

4.

Например, если определяет меру сходства между объектами, то очевидно, что



Мера сходства (различия) задается в числовой форме и указывает степень естественных связей:

- между объектами в группе;

- между объектами и группой объектов;

- между группами объектов.

**5.**

***Примеры функций (***используемых в качестве мер сходства***).***

***Евклидово расстояние*** (простейшая мера).

Для многомерного случая, когда все координаты равнозначны, она имеет следующий вид:



Или рассматривается взвешенное расстояние:



где ‒ весовой коэффициент для *k*-й координаты.



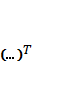
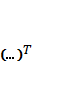
**6.**

***Расстояние Махаланобиса.***

Квадратичное расстояние между объектами записывается в виде



где ‒трансформированный вектор; ‒ обратная ковариационная матрица.



**7.**

***Коэффициент Танимото.***

Танимото предложил отношение подобия:



где ‒ соответственно общее число совпадающих признаков у объектов и .



8.

Проблема разбиения исходных данных на кластеры остается открытой и после выбора меры сходства.

*Качество разбиения* определяется *критерием*, который:

- либо воспроизводить некоторые эвристические соображения (опыт и интуицию исследователя),

- либо основываться на оптимизации показателя качества.

8а.

*Евклидово расстояние* хорошо для эвристического подхода, что связано с естественностью его интерпретации как меры сходства (близости).

Однако поскольку близость двух объектов является относительной мерой их подобия, то вводят порог (параметр).

Порог позволяет установить приемлемые степени сходства при построении кластеров.

9.

Показателем качества является сумма квадратов ошибки:



где *l* ‒ число кластеров; ‒ множество объектов кластера *j*;



‒ вектор его выборочных средних значений.



Критерий определяет совокупное расстояние между объектом и математическим ожиданием кластера, к которому отнесен объект *X*,

Объекты включают в кластер так, чтобы кластеры имели наибольшую возможную «плотность».

10.

***Классификация алгоритмов кластеризации.***

Алгоритмы кластеризации можно разделить на 2 класса:

- *прямые* (конструктивные, «собирательные»);

- *обратные* (оптимизационные, «разделительные»).

В первом случае процесс кластеризации начинается с изолированного объекта и в соответствии с заданным порогом объединяет ближайшие объекты в группы (направление кластеризации снизу вверх).

Во втором ‒ в зависимости от значения (максимума или минимума) критерия объекты распределяются в оптимальные кластеры (направление сверху вниз).

Рассмотрим примеры типичных схем.

12.

**4.2 Алгоритмы кластеризации**

Пусть ‒ исходное множество объектов.



***Алгоритм 1 (число классов неизвестно).***

**Шаг 1.** Первый объект выбирается как представитель первого кластера: , где ‒ центр кластера.



**Шаг 2.** Рассматривают очередной объект *X* и вычисляют его расстояние до всех существующих кластеров:

- *X* включают в , если ;



- *X* не включают в , если ;



- решение не принимается, если *X* попадает в зону неопределенности кластера (см. следующий рис.).

12а.

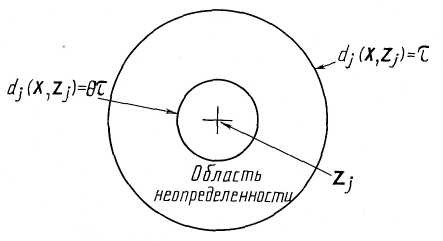


Рис. Зоны принятия решений

13.

**Шаг 3.**

а) когда новый объект *X* включен в кластер , вычисляют его параметры:



*k* ‒ число объектов в классе *X* ‒ -й объект.



и - оценки матожидания и дисперсии.



б) формируют новый кластер , если для любого .



**14.**

**Шаг 4.** Шаги 2, 3 повторяют пока все объекты не распределены по кластерам.

При повторном рассмотрении данных в исходном порядке возможны перераспределения отдельных объектов.

**Шаг 5.** После завершения обучения (когда объекты не изменяют своей классовой принадлежности), полученную систему распознавания можно использовать для классификации большего числа объектов.

К данному моменту зон неопределенности уже не должно быть.

*Замечание.* Все объекты, не попавшие ни в одну из зон, могут, например, включаться в ближайший по минимуму расстояния кластер.

15.

*Недостатки* описанной схемы:

- зависимость результатов от выбора первого центра;

- от порядка рассмотрения объектов;

- от значения порогов.

*Преимущества*:

- позволяет быстро получать приблизительные оценки основных характеристик заданного набора данных;

- является привлекательной с вычислительной точки зрения.

Для понимания геометрии распределения объектов, необходимо проводить эксперименты с различными значениями порога х и исходными точками кластеризации.

16.

Пусть, число классов, на которое необходимо разбить множество объектов , заранее определено.



Предлагается схема, основанная на минимизации суммы квадратов расстояний от всех объектов кластера до его центра:



‒ область -го кластера с центром , полученная на -й итерации.



17.

***Алгоритм 2.***

**Шаг 1.** Произвольным образом выбирают *l* исходных центров кластеров :



**Шаг 2.** На *k*-й итерации распределяют объекты по кластерам в соответствии с правилом:



для всех .



В случае равенства объекты распределяются произвольно.

18.

**Шаг 3.**

Корректируют центры кластеров по формуле:



где ‒ число объектов в кластере .



*Примечание.* Уточненный центр кластера будет минимизировать сумму квадрата расстояний между всеми объектами из .



19.

Центры новых кластеров выбираются таким образом, чтобы минимизировать показатель качества



**Шаг 4.**

Когда , алгоритм заканчивается (условие сходимости).



В противном случае переходим к шагу 2.

*Замечание.* Практическое применение алгоритма также требует проведения численных экспериментов.