**Цель работы**

Реализовать программно в среде MATLAB алгоритм k‑внутригрупповых средних. Для использования алгоритма, в качестве исходных данных, взять «Ирисы Фишера».

**Краткие теоретические сведения**

**Алгоритм k-внутригрупповых средних**

Данный алгоритм минимизирует показатель качества, определённый как сумма квадратов расстояний всех точек, входящих в кластерную область до центра кластера. Это процедура, которую часто называют алгоритмом основанным на k-внутригрупповых средних, состоит из следующих шагов.

**Шаг 1.** Выбирается K исходных центров кластеров Z1(1), Z2(1),…, Zk(1). Этот выбор осуществляется произвольно и обычно в качестве исходных центров используются первые K элементов выборки.

**Шаг 2.** На k-ом шаге итерации заданное множество образцов {X} распределяется по K-кластеров по следующему правилу:

(1)

для всех i = 1,2,…,k, i ≠ j, где Sj(k) – множество образов, входящих в кластер с центром Z­j(k).

В случае равенства (1) решение принимается произвольным образом.

**Шаг 3.** На основе результатов шага 2 определяются новые центры кластеров на k+1 шаге, т.е. Z­j(k+1), j = 1,2,…,K. Исходя из условия, что сумма квадратов расстояний между всеми образами, принадлежащими множеству Sj(k) c новым центром кластера должна быть минимальна. Другими словами: новые центры кластеров Z­j(k+1) выбираются таким образом, чтобы минимизировать показатель (критерий) качества

(2)

Центр Z­j(k+1), обеспечивающий минимизацию показателя качества (2) является выборочным средним, определяемым по множеству Sj(k), следовательно, новые центры кластеров определяются

(3)

Очевидно, что название алгоритма определяется способом, принятым для последовательной коррекции назначения центров кластеров.

**Шаг 4.** Равенство Z­j(k+1) = Z­j(k), j = 1,2,…,K является условием сходимости алгоритма и при его достижении выполнение алгоритма заканчивается. В противном случае надо выбрать K новых центров кластеров и алгоритм повторяется с шага 2.

Качество работы алгоритмов, основанных на вычислении k‑внутригрупповых средних, зависит от числа выбираемых центров кластеров, от геометрических особенностей (структуры) данных.

Получение приемлемых результатов можно ожидать когда данные образуют характерные «гроздья», отстоящие далеко друг от друга. Практическое применение алгоритма потребует проведение экспериментов с целью выбора различных значений параметров K.

**Интерфейс программы**

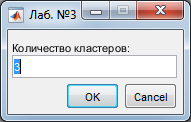


Рисунок 1 — Окно ввода параметров алгоритма.

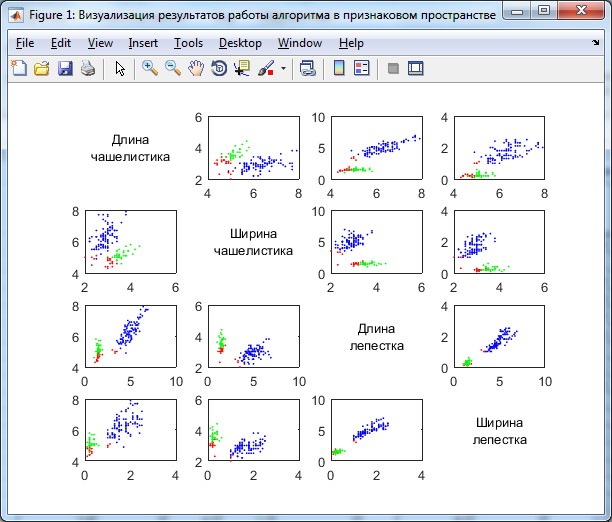


Рисунок 1 — Графическое представление результатов кластеризации.

Результат работы программы также представлен в Excel-файле:

Для каждого экземпляра ириса указывается определённый класс.

Отображаются первоначальные центры кластеров для каждого класса.

Определяется качество работы алгоритма – процент правильно классифицированных экземпляров (только для 3 кластеров).