УДК 004.622

Харченко Дарья Викторовна,
Харченко Степан Александрович,
магистранты кафедры «Вычислительные машины и комплексы»,
ФГБОУ ВО «Ангарский государственный технический университет»,
Истомин Андрей Леонидович,
д.т.н., профессор,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail:a.l.istomin@mail.ru

КЛАСТЕРНЫЙ АНАЛИЗ И ЕГО РЕАЛИЗАЦИЯ В ПАКЕТЕ «STADIA»

Kharchenko D.V., Kharchenko S.A., Istomin A.L.

CLUSTER ANALYSIS AND ITS IMPLEMENTATION IN THE "STADIA" PACKAGE

Аннотация. В статье приведен обзор методов и алгоритмов кластерного анализа и их реализация в пакете «Stadia».

Ключевые слова: кластерный анализ, кластер, аггломеративная стратегия, дивизивная стратегия, функции расстояний, дендрограмма.

Abstract. The review of methods of the cluster analysis.

Keywords: Cluster analysis, cluster, agglomerative strategy, divisive strategy, distance functions, dendrogram.

В настоящее время практически во всех областях человеческой деятельности существует потребность в изучении статистических данных, которые описывают поведение наблюдаемых объектов, событий, процессов или явлений. Одной из наиболее актуальных и востребованных задач анализа данных является задача разбиения множества объектов на сравнительно однородные группы (подмножества), называемые кластерами. Внутри каждой группы должны оказаться элементы максимально «схожие» между собой, а элементы из разных групп должны быть максимально «отличными» друг от друга.

Процедура кластеризации – зависит от степени сходства или не сходства. Такие меры выражаются в виде функций расстояний. Эти расстояния могут определяться в одномерном или многомерном пространстве. Наиболее распространенный способ - вычисление евклидова расстояния между точками i и j в пространстве, когда известны их координаты:

$$d_{E_{ij}} = \left(\sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^2 \right)^{\frac{1}{2}}.$$

Кроме евклидова расстояния, используются и другие меры сходства, называемые метриками или функциями расстояний (линейное расстояние, квадрат евклидова расстояния, обобщенное степенное расстояние Минковского, расстояние Чебышева, расстояние городских кварталов (Манхэттенское расстояние и др.)).

Методы кластерного анализа можно разделить на две группы: иерархи-

ческая (агломеративные и дивизивные методы) и неиерархическая (итеративные методы), которые имеют множество алгоритмов. Применяя различные методы кластерного анализа можно получить множество решений для одних и тех же данных.

В пакете «Stadia» (Statistical Dialogue System) метод кластерного анализа позволяет построить систему классификации объектов или переменных в виде дерева-дендрограммы, а также разбить объекты или переменные на заданное число удаленных друг от друга классов.

Рассмотрим процедуру решения задачи методом кластерного анализа в пакете «Stadia».

Кластерный анализ проводится в несколько этапов:

1. Отбор и преобразование переменных для анализа.

Исходные данные представляются в виде матрицы размером m ×n.

2. Выбор меры расстояния между объектами.

Если исходные данные в виде матрицы размером m ×n, то из меню (рисунок 1) необходимо выбрать метод вычисления расстояния d_{ij} между объектами в многомерном пространстве (метрику).

3. Выбор метода кластеризации.

Агломеративные методы (объединяющие) позволяют создавать дендрограммы. Дивизивные методы (разделяющие) объединяют данные в заданное число кластеров (рисунок 2).

4. Определение числа кластеров.

5. Интерпретация и оценка достоверности кластеров.

Рисунок 1 – Меню выбора метрики расстояний

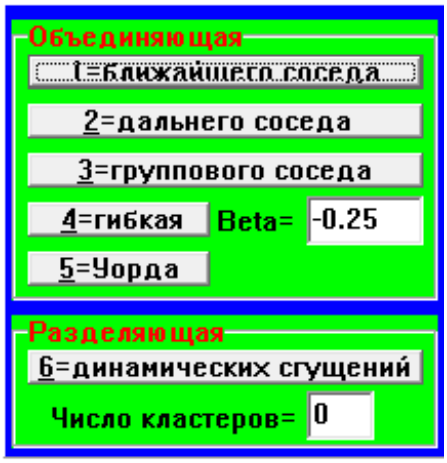

Рисунок 2 – Меню выбора стратегии классификации

ЛИТЕРАТУРА

1. Дюран, Б. Кластерный анализ / Б. Дюран, П. Оделл. – М.: Статистика, 2007. – 128 с.

2. Кулаичев А.П. Методы и средства анализа данных в среде Windows. STADIA 6.0. – М.: Информатика и компьютеры, 1996. – 257 с.