Кластер (Cluster)

В статистике — подмножество объектов статистической совокупности, однородных по своим признакам. В анализе данных и машинном обучении кластер — это область многомерного пространства, расстояние между любыми векторами объектов внутри которой меньше, чем до любого объекта вне кластера. В этом случае векторы объектов образуют явно выделяющиеся «сгустки» в многомерном пространстве признаков.

Понятие кластера вводится и в других предметных областях — астрономии, медицине и биологии, физике, химии, социальной сфере, компьютерной технике и т.д. Но во всех случаях под кластером понимают примерно одно и то же — группу объектов, сходных по своим свойствам, целям и назначению.

Кластер, как объединение однородных объектов, может рассматриваться как самостоятельная единица исследования, обладающая определенным набором свойств. Иными словами, исследуя кластер, мы исследуем не отдельные попавшие в него объекты, а свойства кластера в целом, обобщая их на каждый объект в кластере и на каждый новый объект, который попадет в этот кластер в будущем.

Данный процесс называется содержательной интерпретацией кластера, результатом которой являются правила, зависимости и закономерности, отвечающие на вопросы: в чем сходство объектов в кластере и их отличие от объектов в других кластерах. Если выработан некоторый механизм принятия решений в отношении какого-то объекта в кластере, то этот же механизм может быть применен к любому другому объекту, попавшему в кластер.

Например, если в результате содержательной интерпретации кластера, построенного на данных клиентов банка, обнаружилось, что почти все попавшие в него заемщики являются добросовестными, то можно считать таковым и любого нового клиента, попавшего в этот же кластер.

Следует понимать отличие кластера от класса, хотя оба содержат объекты, близкие по своим свойствам. Классы и их свойства задаются априорно, в то время как кластеры формируются исключительно на основе близости значений признаков объектов, а свойства выясняются в процессе их содержательной интерпретации.

Кластер является основным понятием важного направления в аналитических технологиях — кластерного анализа. Технология обнаружения и формирования кластерных структур с помощью различных аналитических моделей получила название кластеризация.

Впервые термин «кластер» (англ. cluster — гроздь, сгусток, пучок) в контексте описания кластерного анализа был предложен математиком Р. Трионом.

Специальный обработчик кластеризация производит в Loginom кластеризацию объектов на основе алгоритмов k-means и g-means. А в статье «Алгоритмы кластеризации на службе Data Mining» описан целостный взгляд на достижения в области разработки эффективных подходов к кластеризации данных.