Использование человеко-машинных процедур в решении задач кластерного анализа

Шевченко Г.Я. к.т.н., Шумейко О.О. д.т.н., Белозубенко В.С. д.э.н., Исамбаев В.В.

Научный центр Noosphere, Днепровский государственный технический университет

Одним из основных этапов любого исследования является структуризация или кластеризация полученных данных. И визуальное представление такой структуризации играет немаловажную, а иногда и главную роль, особенно когда требуется анализировать многомерные данные, представляющие определенную сложность для многих исследователей-предметников, мало знакомых с основами Data Science. Это связано с чувственным происхождением всего нашего знания и визуализация фактически выполняет функцию средства связи между объектом исследования и исследователем, причем визуализация позволяет увидеть экспериментальные данные в целом, наглядно, при этом иссследователь воспринимает наибольшее количество информации.

Впервые такой акцент на визуальном представлении данных сделал замечательный американский математик и статистик Дж.Тьюки. Настаивая на визуализации, он подчеркивал, что числовые свертки данных делают акцент на ожидаемом, а графические представления – **на неожиданном**. Он же, развивая свой подход, пришел к идее разведочного анализа, который теперь называют Data Mining. Визуализация в анализе данных сегодня применяется все более активно, приводя к смене парадигм в статистике - вместо формализации визуализация, о чем говорили К.Исикава, Г.Тагути и др.

Очень часто, при проведении разведочного анализа, куда можно включить и кластеризацию, используются так называемые таблицы объект-свойство (ТОС). При проведении процедуры кластеризации в большинстве случаев требуется предварительно указать число кластеров, исходя либо из каких-то предположений либо прибегая к специальным вычислениям. Однако такие предположения связаны со значительными издержками - необходимо глубокое знание изучаемого предмета исследований. С другой стороны, вычисление критериев дают неоднозначные результаты - приходится делать выбор между ними, по сути дела достаточно произвольный. В то же время качественное сведение первичных данных ТОС в дву- или трехмерное представление позволяет эффективно решить задачу визуального выявления, по крайней мере, количества кластеров.

Предлагаемые процедуры визуализации основываются на методах многомерного шкалирования и методе главных компонент, позволяющих свести многомерное представление данных к двух- или трехмерным представлениям с сохранением структуры и пропорций, характерных для исследуемой ТОС. Однако зачастую одной визуализации бывает недостаточно для проведения полностью кластерного анализа, поэтому желательно и даже нужно рассматривать сочетание визуализации и кластеризации.

Такого рода соображения, с учетом появления значительного количества исследователей, мало знакомых с техникой кластеризации, и с учетом появления и становления новой парадигмы - автоматизации умственной деятельности, а также другой парадигмы - визуализации вместо формализации приводят нас к необходимости разработки человеко-машинной процедуры решения задач кластеризации - соединения возможностей человека в области визуальной оценки структурированности данных, которые у него представлены достаточно хорошо и вычислительных возможностей компьютера для обработки остальных алгоритмических операций при кластеризации, т.е. получения человеко-машинной процедуры (автоматизация умственной деятельности) кластеризации данных, представленных ТОС. Процедура реализована в виде веб-сервиса, расположенного по адресу: *https://www.sciencehunter.net/Services/Clustering#/visualization*.

В докладе подробно рассматривается такая процедура и приведены примеры решения ряда практических задач кластеризации данных, в том числе известных, для проведения сравнения и подтверждения указанной процедуры.

Доклад на конференции ITMM_2019. Дн-ск, Метал.академия, март 2019г.