АЛГОРИМТ КАЧЕСТВА ОЦЕНКИ КЛАСТЕРИЗАЦИИ

А.В. Быков, М.В. Холманский, А.В. Аксенов
Национальный Исследовательский Томский Политехнический Университет, г. Томск
E-mail: bykov_alexander@bk.ru

Введение

Объектом исследования является алгоритм качества оценки кластеризации.

Целью данной исследовательской работы является разработка приложения определяющего качество кластеризации с использованием алгоритма качества.

Для начала в своей работе хотим дать определение что такое кластеризация? Кластеризация – автоматическое разбиение элементов некоторого множества на группы в зависимости от их схожести (имеющие одинаковые элементы). Слово «кластеризация» имеет множество синонимов основными являются «таксономия», «автоматическая классификация», «обучение без учителя».

Весь процесс кластеризации зависит только от выбранного метода который всегда является итеративным. Он может стать увлекательным процессом и включать множество экспериментов, правильных построенных методов по выбору разнообразных параметров, например, меры расстояния, типа стандартизации переменных, количества кластеров и т.д. Однако эксперименты не должны быть самоцелью - ведь конечной целью кластеризации является получение содержательных сведений о структуре исследуемых данных. А именно получение результатов требующиеся для дальнейшей интерпретации, исследования и изучения свойств и характеристик объектов для возможности точного описания сформированных кластеров.

Также хотим отметить, что в результате применения разных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы, а некоторые методы могут создавать кластеры произвольной формы. Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных), либо предполагать в наборе данных наличие кластеров различного размера. Разные методы кластерного анализа особенно чувствительны к шумам или выбросам, другие - менее. В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма.

Проектирование алгоритма оценки качества.

В своей статье хотим рассмотреть алгоритм оценки качества – это алгоритм Кохонена.

Сеть Кохонена - это наверное одна из основных разновидностей нейронных сетей, которые используют неконтролируемое обучение. При таком обучении обучающее множество состоит лишь из значений входных переменных, в процессе обучения нет сравнивания выходов нейронов с эталонными значениями. Можно сказать, что такая сеть учится понимать структуру данных.

Сеть Кохонена использует следующую модель (рис. 1): сеть состоит из M нейронов, образующих прямоугольную решетку на плоскости — слой.

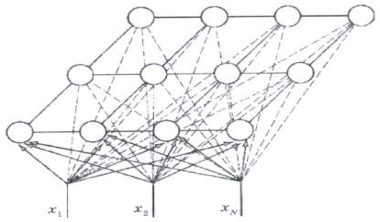

Рис. 1 - Модель сети Кохонена

К нейронам, расположенным в одном слое, представляющем собой двумерную плоскость, подходят нервные волокна, по которым поступает N-мерный входной сигнал. Каждый нейрон характеризуется своим положением в слое и весовым коэффициентом. Положение нейронов, в свою очередь, характеризуется некоторой метрикой и определяется топологией слоя, при которой соседние нейроны во время обучения влияют друг на друга сильнее, чем расположенные дальше. Каждый нейрон образует взвешенную сумму входных сигналов с $w_{ij} > 0$, если синапсы ускоряющие, и $w_{ij} < 0$ - если тормозящие. Наличие связей между нейронами приводит к тому, что при возбуждении одного из них можно вычислить возбуждение остальных нейронов в слое, причем это возбуждение с увеличением расстояния от возбужденного нейрона уменьшается. Поэтому центр возникающей реакции слоя на полученное раздражение соответствует местоположению возбужденного нейрона. Изменение входного обучающего сигнала приводит к максимальному возбуждению другого нейрона и соответственно — к другой реакции слоя. Сеть Кохонена может рассматриваться как дальнейшее развитие LVQ (Learning Vector Quantization). Отличие их состоит в способах обучения.

Сеть Кохонена, в отличие от многослойной нейронной сети, очень проста; она представляет собой два слоя: входной и выходной. Элементы карты располагаются в некотором пространстве, как правило, двумерном.

Сеть Кохонена обучается методом последовательных приближений. В процессе обучения таких сетей на входы подаются данные, но сеть при этом подстраивается не под эталонное значение выхода, а под закономерности во входных данных. Начинается обучение с выбранного случайным образом выходного расположения центров.

В процессе последовательной подачи на вход сети обучающих примеров определяется наиболее схожий нейрон (тот, у которого скалярное произведение весов и поданного на вход вектора минимально). Этот нейрон объявляется победителем и является центром при подстройке весов у соседних нейронов. Такое правило обучения предполагает "соревновательное" обучение с учетом расстояния нейронов от "нейрона-победителя".

Обучение при этом заключается не в минимизации ошибки, а в подстройке весов (внутренних параметров нейронной сети) для наибольшего совпадения с входными данными.

Основной итерационный алгоритм Кохонена последовательно проходит ряд эпох, на каждой из которых обрабатывается один пример из обучающей выборки. Входные сигналы последовательно предъявляются сети, при этом желаемые выходные сигналы не определяются. После предъявления достаточного числа входных векторов синаптические веса сети становятся способны определить кластеры. Веса организуются так, что топологически близкие узлы чувствительны к похожим входным сигналам.

В результате работы алгоритма центр кластера устанавливается в определенной позиции, удовлетворительным образом кластеризующей примеры, для которых данный нейрон является "победителем". В результате обучения сети необходимо определить меру соседства нейронов, т.е. *окрестность* нейрона-победителя, которая представляет собой несколько нейронов, которые окружают нейрон-победитель.

Слой Кохонена состоит из некоторого количества n параллельно действующих линейных элементов. Все они имеют одинаковое число входов m и получают на свои входы один и тот же вектор входных сигналов $x=(x_1,\dots x_m)$

На выходе jго линейного элемента получаем сигнал

$$y_j = w_{j0} + \sum_{i=1}^{m} w_{ji} x_i,$$

где w_{ji} — весовой коэффициент iго входа jго нейрона, w_{j0} — пороговой коэффициент.

После прохождения слоя линейных элементов сигналы посылаются на обработку по правилу «победитель забирает всё»: среди выходных сигналов y_j ищется максимальный; его номер $j_{max} = \operatorname{argmax}_j \{y_j\}$. Окончательно, на выходе сигнал с номером j_{max} равен единице, остальные — нулю. Если максимум одновременно достигается для нескольких j_{max}, то либо принимают все соответствующие сигналы равными единице, либо только первый в списке (по соглашению). «Нейроны Кохонена можно воспринимать как набор электрических лампочек, так что для любого входного вектора загорается одна из них.»

В данной работе рассмотрено только вкратце два решение из поставленных задач, которые будут рассмотрены в полной исследовательской работе:

1. Сравнение индексов оценки качества кластеризации.
2. Разработка модели качества кластеризации.

Также хочется сказать, что было рассмотрен основной и очень распространённый алгоритм качества который и будет основным алгоритмом лежащий в проектирование предметной области и написание программы для определения качества кластеризации с использованием разных алгоритмов качества.

Список использованной литературы:

1. Руденко О.Г., Бодянский Е.В. Искусственные нейронные сети – Харьков, 2005.
2. Котов А., Красильников Н. Кластеризация данных. 2006.
3. Иерархический алгоритм [Электронный ресурс]. – Режим доступа: http://www/csee/umbc/edu/nicolas/clustering/p264-jain.pdf , свободный