МЕТОД КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ НА ОСНОВЕ МОДИФИЦИРОВАННОГО АЛГОРИТМА ФУНКЦИОНИРОВАНИЯ КАРТ КОХОНЕНА.

Герасимова Н.И.
Научный руководитель –Аксенов С.В., к.т.н., доцент
Томский Политехнический Университет, Институт кибернетики
nig1@tpu.ru

В настоящее время объемы массивов данных возрастают, а для их обработки требуются усовершенствованные методы интеллектуального анализа данных. Кластерный анализ является одним из важных и известных этапов интеллектуального анализа данных. Задачей кластеризации является обнаружение во множестве исходных данных таких групп объектов, что объекты из одной группы должны быть как можно больше похожи друг на друга, а объекты из разных групп как можно больше отличаться друг от друга. Такие группы объектов называются кластерами. [1]

За время существования кластерного анализа было разработано большое количество различных алгоритмов и методов. До сих пор алгоритмы дорабатываются и изменяются в зависимости от потребностей области и задач применения кластеризации. В связи с этим актуальна потребность в проведении исследования работы стандартных алгоритмов кластеризации, а также их модификаций с целью обнаружения особенностей их работы, выявления достоинств и недостатков, а также повышения эффективности их работы в зависимости от специфики обрабатываемых данных.

Самоорганизующиеся карты Кохонена (самоорганизующиеся карты признаков) – это нейронная однослойная сеть прямого распространения, в которой используется обучение без учителя. Такие сети отличаются от нейронных сетей с обучением с учителем тем, что их обучающая выборка состоит только из значений входных сигналов, а значения на выходе формируются самостоятельно и не сравниваются со значениями-эталонами.

При реализации алгоритма кластеризации на основе самоорганизующихся карт Кохонена количество нейронов в сети задается заранее, и перед началом обучения случайным образом инициализируются весовые коэффициенты всех нейронов.

Обучение самоорганизующихся карт Кохонена производится методом последовательных приближений с использованием итераций, в которых корректируются нейроны-векторы. На каждой итерации из исходной выборки случайным образом выбирается один вектор, а затем происходит поиск наиболее похожего на него вектора коэффициентов нейронов. Нейрон, наиболее схожий с входным вектором, объявляется нейроном-победителем. Мера схожести между векторами рассчитывается при помощи евклидова расстояния. Таким образом, для i-го нейрона-победителя выполняется равенство:

$$d(x, w_i) = min_{1 \le j \le n}\, d\,(x, w_j),$$

где n – количество нейронов, а $d(x, w_i)$ – расстояние между векторами xи w. После этого корректируются все веса нейронной сети, таким образом, что вектор-нейрон-победитель и его соседние вектор в сетке сдвигаются по направлению к входному вектору, модифицируется, как и нейрон-победитель, так и его соседи (но в меньшей мере). [2]

Весовые коэффициенты нейронов-победителей и нейронов, лежащих в его окрестности, изменяются по следующей формуле:

$$w_i(k+1) = w_i(k) + \eta_i(k)[x - w_i(k)],$$

где $\eta_i(k)$ – коэффициент обучения i-го нейрона, принадлежащего окрестности $S_w(k)$ в k-й момент времени. Таким образом, изменение значения весового коэффициента i-го нейрона тем меньше, чем дальше этот нейрон расположен от нейрона-победителя. Если нейрон не принадлежит окрестности $S_w(k)$, то его весовой коэффициент изменяться не будет. Выполнение алгоритма будет происходить до тех пор, пока выходные значения не появятся на карте признаков. После окончательного размещения нейронов отображается полученная карта. [3]

Кластеризация с помощью самоорганизующихся карт Кохонена накладывает ограничение на форму кластера, в частности с помощью данного алгоритма нельзя распознавать кластеры произвольной и сложной формы. Также при использовании алгоритма Кохонена нужно строго задавать количество кластеров. Для снятия данного ограничения оригинальный алгоритм подвергся модификации так, чтобы он смог распознавать кластеры любой формы при неизвестном заранее числе кластеров. Данная модификация строится на анализе межкластерных расстояний для групп векторов, которые связаны с соседними нейронами в самоорганизующейся карте Кохонена.

На первом этапе алгоритма проводится кластеризация данных с помощью алгоритма Кохонена. Затем производится процесс обнаружения «мёртвых» кластеров, то есть таких, которые не содержат нейронов. Они удаляются из списка анализируемых кластеров. После этого для

каждого кластера производится расчет среднего внутрикластерного расстояния по формуле:

$$d = \frac{1}{n-1}\sum_{i}^{n}\sum_{j}^{n} d((x_i, x_j)),$$

где n – количество точек в кластере, $d(x_i, x_j)$ – евклидово расстояние между точками x_i и x_j, п

р Затем для каждого кластера производится поиск соседних кластеров при помощи полученных самоорганизующихся карт признаков. Благодаря таким картам соседние кластеры находятся быстрее, так как не затрачивается время на обход всех кластеров, а по расположению кластеров в соседних узлах карты определяются «кластеры-соседи».

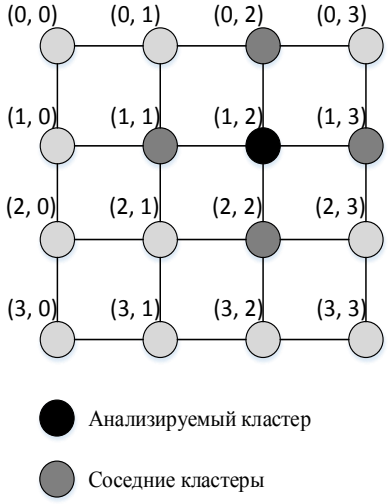

Рисунок 8 - Пример нахождение соседних кластеров в самоорганизующейся карте признаков (сетка 4x4)

После этого происходит анализ расстояний между двумя кластерами. Если расстояние между точками двух кластеров меньше, чем большое из внутрикластерных расстояний двух кластеров, то эти точки считаются граничными точками двух кластеров, а кластеры подходят для объединения в один супекластер.

Минимальное число граничных точек задается количеством либо процентом в зависимости от того, какое из этих значений минимальное, что определяется после определения и сравнения количества точек в двух анализируемых кластерах. В разработанном программном обеспечении минимальное число граничных точек равно двум, а минимальный процент – 10.

Для оценки результатов кластеризации с помощью оригинального и модифицированного

методов использовались индексы оценки качества кластеризации. В качестве основной метрики был выбран индекс силуэта.

Индекс силуэта сравнивает средние значения расстояний между элементами в пределах одного кластера и средние значений расстояний от элементов одних кластеров до элементов других кластеров. Этот индекс использует как критерий компактности, так и критерий отделимости для каждого объекта кластерной структуры. В итоге вычислений будет найден параметр, называемый среднекластерным «силуэтом», значение этого параметра показывает, насколько плотно сгруппированы точки внутри одного конкретного кластера. А также «силуэт» всей кластерной структуры, который является средним значением среднекластерных «силуэтов» и показывает, насколько точно была проведена кластеризация. [4]

Для эффективной и объективной оценки качества полученных кластерных структур результаты полученной кластеризации также были оценены с помощью индекса Данна и RS индекса.

Работа алгоритма тестировалась на наборах данных двух видов: данные, сгенерированные случайным образом и «ирисы Фишера». После применения модифицированного алгоритма индексы качества показали более высокие значения, чем после применения оригинального метода, что свидетельствует о качественном и эффективном проведении модификации алгоритма кластеризации.

Список литературы

1. Мандель, И. Д. Кластерный анализ / И.Д. Мандель. — М.: Финансы и статистика, 1988. — 176 с.

2. Нейронные сети Кохонена [Электронный ресурс] / Портал искусственного интеллекта [2015]. Режим доступа: h

t 3. Самоорганизующиеся карты Кохонена - математический аппарат [Электронный ресурс] / Labs. Технологии анализа данных [2015]. Режим доступа:

w

w

a 4. Clustering Indices [Электронный ресурс] / The Comprehensive R Archive Network [2015]. Режим

p

o

c

t

r

y

h

a

u

a

r

t

i

c

l

e

s

n

e