Анисимова Э.С. ©
Ассистент, кафедра информатики и дискретной математики
Елабужский институт Казанского федерального университета

САМООРГАНИЗУЮЩИЕСЯ КАРТЫ КОХОНЕНА В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ

Аннотация

Статья рассматривает одну из технологий кластеризации – самоорганизующиеся карты Кохонена. Представлены алгоритм работы, структура сети.

Ключевые слова: самоорганизующиеся карты, кластеризация, интеллектуальный анализ данных.

Keywords: self-organizing maps, clustering, data mining.

Самоорганизующаяся карта Кохонена — соревновательная нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Идея сети предложена финским учёным Теуво Кохоненом. Является методом проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего, двумерное), применяется также для решения задач моделирования, прогнозирования и др. В основе идеи сети Кохонена лежит аналогия со свойствами человеческого мозга. Кора головного мозга человека представляет собой плоский лист и свернута складками. Она обладает определенными топологическими свойствами (участки, ответственные за близкие части тела, примыкают друг к другу и все изображение человеческого тела отображается на эту двумерную поверхность).

Структура сети

Сеть Кохонена, в отличие от многослойной нейронной сети, очень проста; она представляет собой два слоя: входной и выходной. Ее также называют самоорганизующей картой.

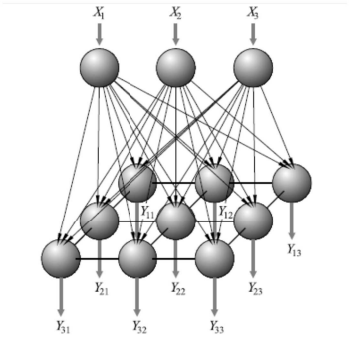

Рис. 1. Самоорганизующаяся карта Кохонена

SOM (*Self-organizing map*) подразумевает использование упорядоченной структуры нейронов. Обычно используются одно и двумерные сетки. При этом каждый нейрон представляет собой n-мерный вектор-столбец $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$, где n определяется размерностью исходного пространства (размерностью входных векторов). При этом, как было сказано выше, нейроны также взаимодействуют друг с другом. Величина этого взаимодействия определяется расстоянием между нейронами на карте.

Алгоритм работы сети

Пусть t— номер итерации (инициализация соответствует номеру 0).

- **Инициализация**

Наиболее распространены три способа задания первоначальных весов узлов:

- Задание всех координат случайными числами.
- Присваивание вектору веса значение случайного наблюдения из входных данных.

- Выбор векторов веса из линейного пространства, натянутого на главные компоненты набора входных данных.

- **Цикл**
- Выбрать произвольное наблюдение *x(t)* из множества входных данных.
- Найти расстояния от него до векторов веса всех узлов карты и определить ближайший по весу узел $M_c(t)$. Это — BMU или Winner. Условие на $M_c(t)$: $\|x(t) - w_c(t)\| \le \|x(t) - w_i(t)\|$, для любого $w_i(t)$, где $w_i(t)$ — вектор веса узла $M_i(t)$. Если находится несколько узлов, удовлетворяющих условию, BMU выбирается случайным образом среди них.
- Определить с помощью функции h(функции соседства) соседей M_c и изменение их векторов веса.

Часто в качестве функции соседства используется гауссовская функция:

$$h_{ci}(t) = \alpha(t) \cdot \exp(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)})$$

где $0 < \alpha(t) < 1$ - обучающий сомножитель, монотонно убывающий с каждой последующей итерацией (то есть определяющий приближение значения векторов веса BMU и его соседей к наблюдению; чем больше шаг, тем меньше уточнение); r_i, r_c - координаты узлов $M_i(t)$ и $M_c(t)$ на карте; $\sigma(t)$ - сомножитель, уменьшающий количество соседей с итерациями, монотонно убывает.

Более простой способ задания функции соседства: $h_{ci}(t) = \alpha(t)$,

если $M_i(t)$ находится в окрестности $M_c(t)$ заранее заданного аналитиком радиуса, и 0 в противном случае. Функция *h(t)* равна *α(t)* для BMU и уменьшается с удалением от BMU.

- Изменить вектор веса по формуле: $w_i(t) = w_i(t-1) + h_{ci}(t) \cdot (x(t) - w_i(t-1))$

- **Вычисление ошибки карты**

Например, как среднее арифметическое расстояний между наблюдениями и векторами веса соответствующих им BMU:

$\frac{1}{N}\sum_{i=1}^{N}\|x_i - w_c\|$, где N - количество элементов набора входных данных.

Раскраска, порожденная отдельными компонентами

При данном методе отрисовки полученную карту можно представить в виде слоеного пирога, каждый слой которого представляет собой раскраску, порожденную одной из компонент исходных данных. Полученный набор раскрасок может использоваться для анализа закономерностей, имеющихся между компонентами набора данных. После формирования карты мы получаем набор узлов, который можно отобразить в виде двумерной картинки. При этом каждому узлу карты можно поставить в соответствие участок на рисунке, четырех или шестиугольный, координаты которого определяются координатами соответствующего узла в решетке. Теперь для визуализации осталось только определить цвет ячеек этой картинки. Для этого и используются значения компонент. Самый простой вариант – использование градаций серого. В этом случае ячейки, соответствующие узлам карты, в которые попали элементы с минимальными значениями компонента или не попало вообще ни одной записи, будут изображены черным цветом, а ячейки, в которые попали записи с максимальными значениями такого компонента, будут соответствовать ячейки белого цвета.

Полученные раскраски в совокупности образуют атлас, отображающий расположение компонент, связи между ними, а также относительное расположение различных значений компонент.

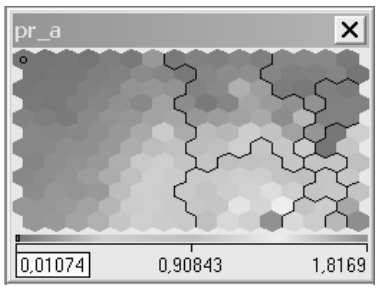

Рис. 2. **Пример карты Кохонена**

Отображение кластеров

Кластером будет являться группа векторов, расстояние между которыми внутри этой группы меньше, чем расстояние до соседних групп. Структура кластеров при использовании алгоритма SOM может быть отображена путем визуализации расстояния между опорными векторами (весовыми коэффициентами нейронов).

Заключение

Основное отличие сетей Кохонена от других моделей состоит в наглядности и удобстве использования. Эти сети позволяют упростить многомерную структуру, их можно считать одним из методов проецирования многомерного пространства в пространство с более низкой размерностью.

Литература

1. E.S. Anisimova – Fractals and digital steganography // Сборник научных трудов Sworld. – 2014. – Т. 6. № 1. – С. 69-71.
2. Э.С. Анисимова – Определение кредитоспособности физического лица в аналитическом пакете Deductor (BaseGroup) // Сборник научных трудов Sworld. – 2014. – Т. 23. № 2. С. – 78-81.
3. А.Ф. Филипов, Э.С. Анисимова – Калькулятор для работы с комплексными числами // Сборник научных трудов Sworld. – 2014. – Т. 29. №2. – С. 47-50.
4. Д.С. Тимофеев, Э.С. Анисимова – Разработка электронного образовательного ресурса на площадке «Тулпар» системы дистанционного обучения КФУ// Сборник научных трудов Sworld. – 2014. – Т.7. №2. –С.80-83.
5. Э.С. Анисимова – Сжатие изображений с помощью квадратичных кривых Безье // Естественные и математические науки в современном мире. – 2014. – № 14. – С. 42-46.
6. Э.С. Анисимова – Формирование математической компетентности студентов психолого-педагогического направления // Сборник научных трудов Sworld. – 2013. – Т. 19. № 4. – С. 56-58.
7. Э.С. Анисимова – Фрактальное кодирование изображений // Сборник научных трудов Sworld. – 2013. – Т. 4. № 3. – С. 79-81.
8. Э.С. Анисимова – Идентификация онлайн-подписи с помощью оконного преобразования Фурье и радиального базиса // Компьютерные исследования и моделирование. – 2014. – Т. 6. № 3. – С. 357-364.