

Нейросетевой кластерный анализ в среде NeuroXL Clusterizer

Осипов Геннадий Сергеевич

Сахалинский государственный университет

д.т.н., заведующий кафедрой Информатики

Вашакидзе Нателла Семеновна

Сахалинский государственный университет

доцент кафедры Информатики

Аннотация

Сформулированы методологические аспекты использования самоорганизующихся сетей Кохонена для решения задачи разделения множества объектов на кластеры. Предложено в качестве инструментария для проведения исследования использовать надстройку над *Excel NeuroXL Clusterizer* компании *OLSOFT LLC*. Рассмотрен практический пример кластеризации совокупности стран по набору показателей.

Ключевые слова: кластерный анализ, нейронная сеть.

Neural network cluster analysis in NeuroXL Clusterizer environment

Osipov Gennadij Sergeevich

Sakhalin State University

Doctor of technical Sciences, Head of the Department of Computer Science

Vashakidze Natella Semenovna

Sakhalin State University.

Associate Professor of Department of Computer Science

Abstract

Methodological aspects of the use of self-organizing Kohonen networks for solving the problem of dividing a set of objects into clusters are formulated. It was proposed to use the add-in for Excel NeuroXL Clusterizer of Olsoft LLC as a tool for conducting the research. A practical example of clustering of a set of countries by a set of indicators is considered.

Keywords Cluster analysis, neural network

Введение

Современные достижения в области искусственного интеллекта во многом базируются на использовании искусственных нейронных сетей. Особое положение здесь занимают самоорганизующиеся нейронные сети, или сети (карты, решетки) Кохонена, которые способны менять свою топологию подстраиваясь под структуру поступающих на них исходных

данных. По сути, это сети, которые «обучаются без учителя», т.е. являются самоорганизующимися.

Одной из наиболее важных и трудоемких задач интеллектуальной обработки информации является проблема разделения исследуемых объектов на группы, классы или кластеры (в многопараметрической постановке) для их последующей классификации и дальнейших исследований. Разделение на классы позволяет в перспективе идентифицировать новый объект и применить для его изучения отработанные методы исследований.

Таким образом данная работа имеет целью отработать основные методологические аспекты нейросетевого подхода к решению задач кластерного анализа. Исследования являются продолжением работ, представленных в [1, 2].

Постановка задачи

Имеется множество систем (объектов, ситуаций), характеризуемых набором параметров. Требуется разбить исходное множество на непересекающиеся подмножества, которые будут содержать в известном смысле близкие друг-другу системы.

Количество подмножеств (кластеров) и степень «компактности» каждого из них могут задаваться пользователем или выбираться автоматически в соответствии с используемым алгоритмом кластеризации.

Инструментарий

Задачи кластеризации и классификации достаточно часто возникают и решаются на практике стандартными «жесткими» алгоритмами, которые не допускают возможности адаптации и своей самоорганизации.

Наибольших успехов отрасль искусственного интеллекта, связанная с проблемой синергетической самоорганизации, достигла благодаря использованию самоорганизующихся сетей (карт, решеток) Кохоненеа, которые относятся к классу искусственных нейронных сетей, функционирующих «без учителя» и способных самостоятельно за счет изменения собственной пространственной топологии эффективно решать задачи разделения исходного множества данных на классы (кластеры, сгустки).

В качестве инструментария для проведения разработки выбрана надстройка над *Excel NeuroXL Clusterizer* [3], которая позволяет пользователю работать в знакомой программной среде и не требует дополнительных знаний из теории нейронных сетей.

Практическая реализация

Решим задачу, сформулированную в [3]. Множество стран, характеризуется унифицированным набором характеристик (параметров – A, B, C, D, E, F), представленных на рисунке 1.

Страна	A	B	C	D	E	F
Алжир	3,9	31193917	1,74	2381740	0	5550
Аргентина	-3	36955182	1,16	2736690	30200	17000
Австралия	4,3	19169083	1,02	7617930	68920	21070
Бразилия	0,8	172860370	0,94	8456510	55455	28000
Канада	3,6	31281092	1,02	9220970	755170	7100
Китай	7	1261832482	0,9	9326410	270550	498720
Индия	5,5	1014003817	1,58	2973190	314400	480000
Индонезия	0	22484210	1,63	1826440	93000	45970
Иран	1	65619636	0,83	1636000	12000	94000
Казахстан	1,7	16733227	-0,05	2669800	47500	22000
Ливия	2	5115450	2,42	1759540	0	4700
Мексика	3,7	100349766	1,53	1923040	49510	61000
Монголия	3,5	2650952	1,54	1565000	0	800
Нигерия	2	10075511	2,75	1266700	300	660
Перу	2,4	27012899	1,75	1280000	5220	12800
Россия	3,2	146001176	-0,38	16995800	79400	40000
Саудовская Аравия	1,6	22023508	3,28	1960582	0	4350
Судан	3	35079814	2,84	2376000	129810	19460
США	4,1	275562673	0,91	9158960	470131	207000

Рис.1 Структура исходных данных в MS Excel

На рисунке 2 представлено заполненное диалоговое окно используемой надстройки, в котором указан диапазон исходных и выходных данных, а также априори определено количество кластеров, на которые необходимо разбить исходной множество стран.

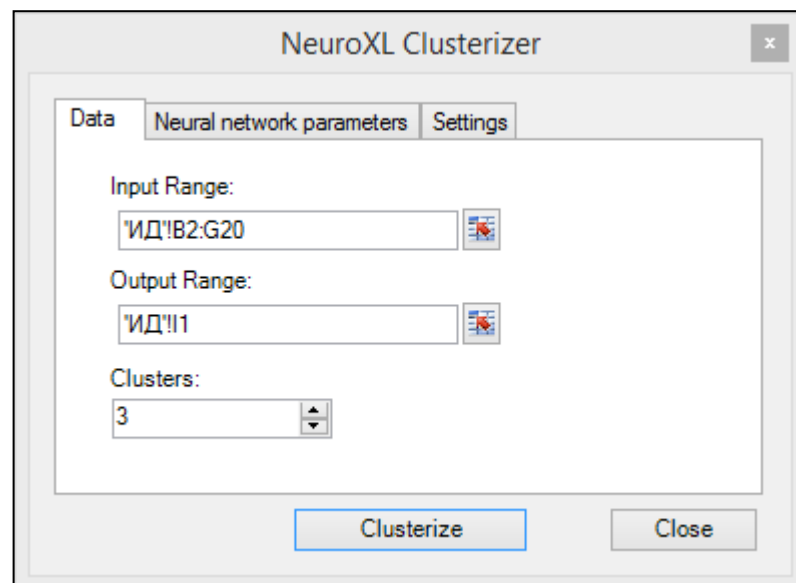


Рис. 2 Заполненная закладка по исходным данным

Закладки по настройке параметров нейронной сети можно (рис. 3) оставить без изменений (согласившись с «рекомендациями» программного средства по умолчанию)

Рис. 3 Параметры настройки

После выполнения кластеризации множество стран автоматически отобразится в *Excel* с учетом их принадлежности к тому или иному кластеру (рис. 4), которые нумеруются и различаются цветом.

Страна	A	B	C	D	E	F	Кластеры
Алжир	3,9	31193917	1,74	2381740	0	5550	1
Аргентина	-3	36955182	1,16	2736690	30200	17000	1
Австралия	4,3	19169083	1,02	7617930	68920	21070	1
Иран	1	65619636	0,83	1636000	12000	94000	1
Казахстан	1,7	16733227	-0,05	2669800	47500	22000	1
Мексика	3,7	100349766	1,53	1923040	49510	61000	1
Монголия	3,5	2650952	1,54	1565000	0	800	1
Нигерия	2	10075511	2,75	1266700	300	660	1
Перу	2,4	27012899	1,75	1280000	5220	12800	1
Россия	3,2	146001176	-0,38	16995800	79400	40000	1
Судан	3	35079814	2,84	2376000	129810	19460	1
США	4,1	275562673	0,91	9158960	470131	207000	2
Бразилия	0,8	172860370	0,94	8456510	55455	28000	2
Канада	3,6	31281092	1,02	9220970	755170	7100	2
Ливия	2	5115450	2,42	1759540	0	4700	2
Саудовская	1,6	22023508	3,28	1960582	0	4350	3
Китай	7	1261832482	0,9	9326410	270550	498720	3
Индия	5,5	1014003817	1,58	2973190	314400	480000	3
Индонезия	0	22484210	1,63	1826440	93000	45970	3

Рис. 4 Преобразование исходных данных

Программа автоматически формирует и размещает на рабочем листе *Excel* информацию о весах и профилях кластеров (рис. 5 и 6).

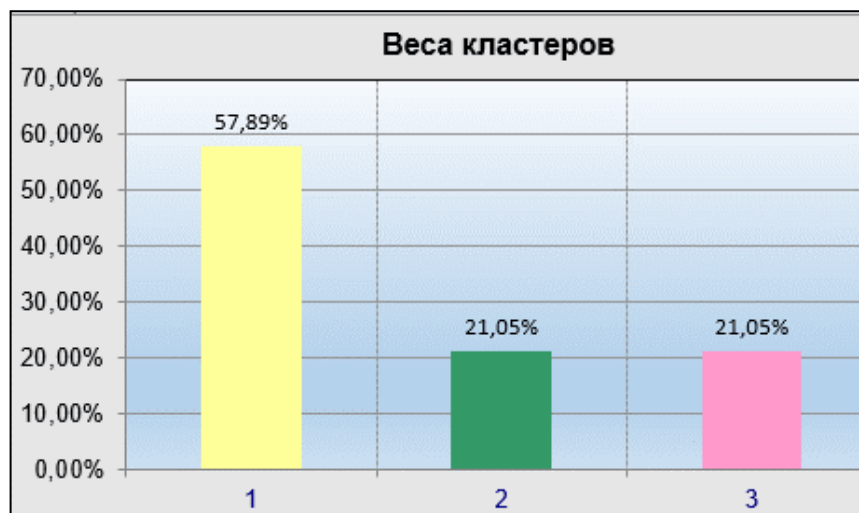


Рис.5 Процент количества объектов в кластерах



Рис. 6 Параметры кластеров

Кроме того, дополнительно выводится информация о минимальном, максимальном и среднем значении параметров в кластерах.

Выводы

Проведенное исследование позволяет сделать следующие выводы:

1. Проблема классификации и кластеризации объектов достаточно эффективно решается с помощью самоорганизующихся искусственных нейронных сетей.

2. Применение *NeuroXL Clusterizer* позволяет упростить решение задачи за счет того, что работа выполняется непосредственно в среде *MS Excel* и, кроме того, от пользователя не требуются профессиональные знания по используемым нейронным сетям.

Библиографический список

1. Technical analysis software for Microsoft Excel. URL:

-
- <http://www.analyzerxl.com/> (дата обращения: 13.08.2017).
2. Осипов Г.С., Вашакидзе Н.С., Филиппова Г.В. Основы прогнозирования финансовых временных рядов на базе NeuroXL Predictor // Постулат. 2017. № 7. С. 26.
 3. NeuroXL Clusterizer. URL: <http://neuroxl.com/products/excel-cluster-analysis-software/neuroxl-clusterizer.htm> (дата обращения: 13.08.2017).