УДК 004

ИСПОЛЬЗОВАНИЕ MATLAB ДЛЯ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ

Суслова Екатерина Владимировна

магистрант
Мордовский государственный университет им. Н.П. Огарёва, Саранск

author@apriori-journal.ru

Аннотация. В статье рассматриваются методы кластерного анализа, а так же сложности применения кластерного анализа данных. Обзор инструментов для программной реализации алгоритмов кластерного анализа.

Ключевые слова: кластерный анализ; нейронная сеть; MATLAB; Statistics Toolbox; Neural Network Toolbox.

USE OF MATLAB FOR THE CLUSTER ANALYSIS OF DATA

Suslova Ekaterina Vladimirovna

undergraduate
Mordovian state university of N.P. Ogaryov, Saransk

Abstract. In article methods of the cluster analysis, and also difficulties of application of the cluster analysis of data are considered. The review of tools for program realization of algorithms of the cluster analysis.

Key words: cluster analysis; neural network; MATLAB; Statistics Toolbox; Neural Network Toolbox.

При анализе и прогнозировании социально-экономических явлений довольно часто сталкиваются с многомерностью их описания. Методы многомерного анализа – количественный инструмент исследования социально-экономических процессов, описываемых большим числом характеристик. К ним относятся кластерный анализ, таксономия, распознавание образов, факторный анализ [1].

Спектр применений кластерного анализа очень широк. Кластерный анализ нашел применение во многих аспектах человеческой жизни.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить «структуру данных».

На сегодняшний день разработано более сотни различных алгоритмов кластеризации. Все они могут быть разделены на иерархические и неиерархические.

Суть иерархической кластеризации (таксономии) состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие. Иерархические методы могут быть агломеративными (объединительными) и дивизивными.

Агломеративная кластеризация начинается с каждого объекта в отдельном кластере. Кластеры объединяют, группируя объекты каждый раз во все более и более крупные кластеры. Этот процесс продолжают до тех пор, пока все объекты не станут членами одного единственного кластера.

Разделяющая, или дивизивная, кластеризация начинается со всех объектов, сгруппированных в единственном кластере. Кластеры делят (расщепляют) до тех пор, пока каждый объект не окажется в отдельном кластере.

Иерархические методы кластерного анализа используются при небольших объемах наборов данных. Преимуществом таких методов кластеризации является их наглядность.

Наиболее распространен среди неиерархических методов алгоритм *k*-средних, также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм *k*-средних строит *k* кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм *k*-средних, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа *k* может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Преимуществом алгоритма являются понятность, прозрачность алгоритма, быстрота и простота реализации. К его недостаткам можно отнести неопределенность выбора начальных центров кластеров, а также то, что число кластеров должно быть задано изначально, что может потребовать некоторой априорной информации об исходных данных. К недостаткам, также относится и то, что алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Помимо всего, алгоритм может медленно работать на больших базах данных.

Другим распространенным методом решения задачи кластеризации данных является применение самоорганизующихся нейронных сетей Кохонена, которые можно обучить на имеющемся наборе данных [2].

У данного подхода есть следующие особенности:

- Искусственные нейронные сети легко работают в распределенных системах с большой параллелизацией в силу своей природы.
- НС оперируют числами, поэтому они могут проводить разбиение на кластеры только для объектов с численными векторами характеристик.

– Нейронные сети менее чувствительны к зашумленным данным.

Нейронные сети Кохонена – класс нейронных сетей, основным элементом которых является слой Кохонена, состоящий из адаптивных линейных сумматоров. Как правило, выходные сигналы слоя Кохонена обрабатываются по правилу «победитель забирает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль. Например, в [3] рассматривалась разработка методики и алгоритмов идентификации отклонений от нормативов параметров качества электроэнергии в системах электроснабжения.

Этап анализа результатов кластеризации подразумевает решение следующих вопросов: не является ли полученное разбиение на кластеры случайным; является ли разбиение надежным и стабильным на подвыборках данных; существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации; можно ли интерпретировать полученные результаты кластеризации.

Существует ряд сложностей, которые следует продумать перед проведением кластеризации.

а) Сложность выбора характеристик, на основе которых проводится кластеризация. Необдуманный выбор приводит к неадекватному разбиению на кластеры и, как следствие, – к неверному решению задачи.

б) Сложность выбора метода кластеризации. Этот выбор требует неплохого знания методов и предпосылок их использования. Чтобы проверить эффективность конкретного метода в определенной предметной области, целесообразно применить следующую процедуру: рассматривают несколько априори различных между собой групп и перемешивают их представителей между собой случайным образом. Далее проводится кластеризация для восстановления исходного разбиения на кластеры. Доля совпадений объектов в выяв-

ленных и исходных группах является показателем эффективности работы метода.

в) Проблема выбора числа кластеров. Если нет никаких сведений относительно возможного числа кластеров, необходимо провести ряд экспериментов и, в результате перебора различного числа кластеров, выбрать оптимальное их число.

г) Проблема интерпретации результатов кластеризации. Форма кластеров в большинстве случаев определяется выбором метода объединения. Однако следует учитывать, что конкретные методы стремятся создавать кластеры определенных форм, даже если в исследуемом наборе данных кластеров на самом деле нет.

Подводя итог всему вышесказанному, можно отметить, что возможность обучения по заданным обучающим выборкам и работа с «зашумленными» данными и в условиях неполноты информации, является несомненным плюсом нейронных сетей. Однако нейросеть – «черный ящик», т.е. практически невозможно извлечь знания, полученные в ходе обучения нейросети. И наконец, самоорганизующиеся карты работают по алгоритму «обучение без учителя», что позволяет применять их для изучения практически любого набора данных, а полученное в ходе обучения самоорганизующейся карты знание доступно для изучения, благодаря чему можно выявлять скрытые закономерности.

Программная реализация алгоритмов кластерного анализа широко представлена в различных программных продуктах.

Одними из наиболее распространенных продуктов, где применяются статистические законы, являются STATISTICA, SPSS Statistics Base, а также MATLAB.

Несмотря на большое количество разнообразных приложений для кластеризации данных, их общим недостатком является ограниченный набор методов кластеризации и возможностей настройки параметров. Данную проблему с успехом решает среда математического моделиро-

вания MATLAB. Помимо встроенных функций кластеризации, пользователь имеет возможность реализовать собственный алгоритм с помощью высокоуровневого объектно-ориентированного языка программирования MATLAB.

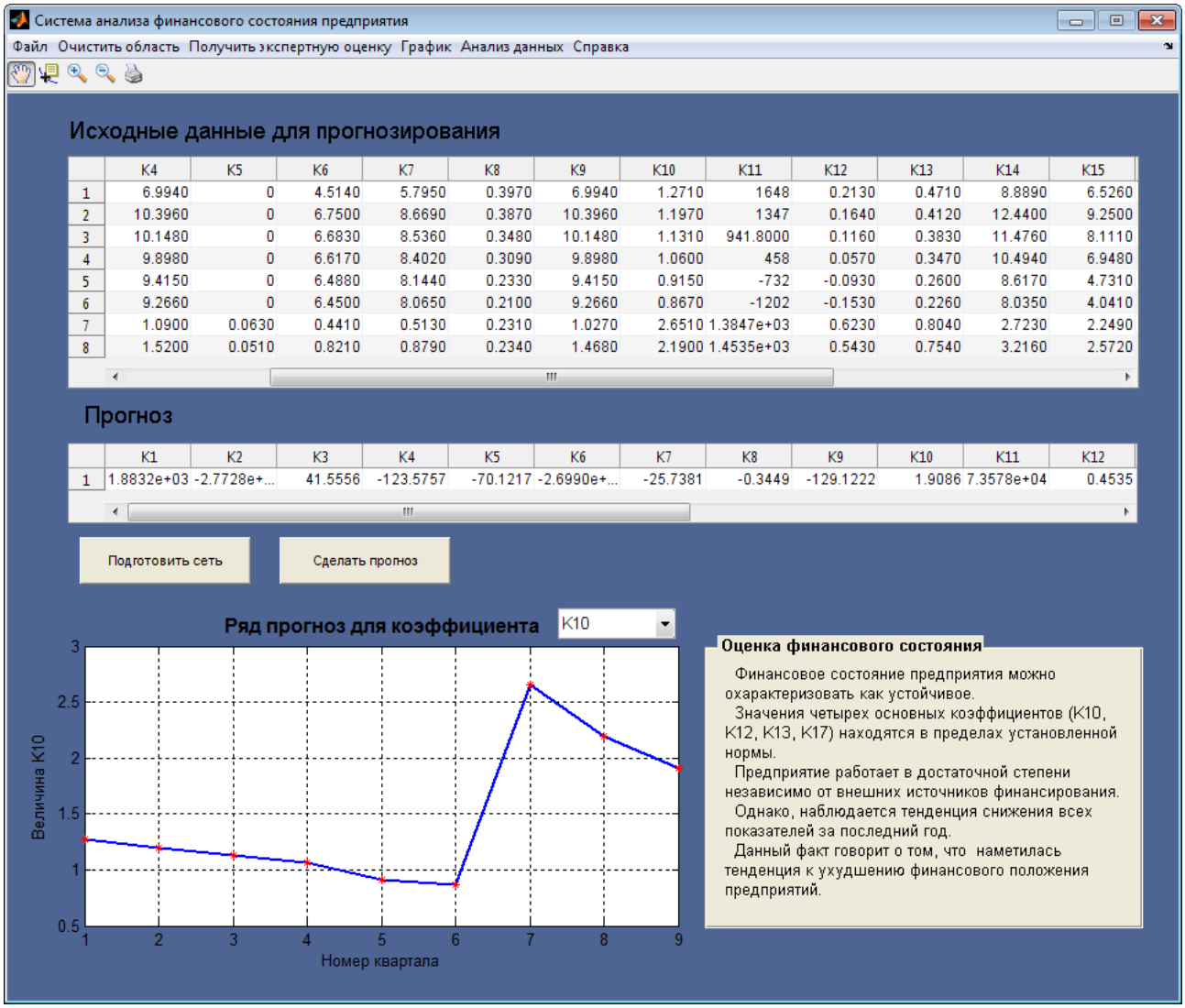
Возможности MATLAB весьма обширны, а по скорости выполнения задач система нередко превосходит своих конкурентов. Она применима для расчетов практически в любой области науки и техники. Этому способствует не только расширенный набор матричных и иных операций и функций, но и наличие пакетов расширения (toolbox, Simulink), специально предназначенных для решения задач блочного моделирования динамических систем и устройств, а также десятков других пакетов расширений, в том числе, статистический пакет – Statistics Toolbox и пакет моделирования нейронных сетей – Neural Network Toolbox [4].

Среда математического моделирования MATLAB находит широкое применение как в научной, так и в учебной деятельности. Например, в [5] основы теории управления сопровождаются программными кодами для системы MATLAB. В работе [6] рассматриваются аспекты обучения программированию на языке С и в системе MATLAB. Следует отметить, что многие конструкции в MATLAB имеют сходство с языком С [6].

Помимо данных функции в качестве инструмента кластеризации данных можно использовать нейронные сети. Система MATLAB включает пакет Neural Network Toolbox™ (NNT), являющийся одним из наиболее гибких и распространенных инструментов создания нейронных сетей. Пакет прикладных программ NNT обеспечивает эффективную поддержку проектирования, обучения и моделирования множества известных сетевых парадигм, от базовых моделей персептрона до самых современных ассоциативных и самоорганизующихся сетей.

Все это, в сочетании с возможностью полностью контролировать процесс создания и обучения сети, предоставляет необходимые возможности для решения задач кластеризации данных.

Для решения задачи кластерного анализа исходных данных для системы анализа финансового состояния предприятий было создано приложение в системе MATLAB на основе утилиты guide. Вид этого приложения приведен на рис. 1.



	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14	K15
1	6.9940	0	4.5140	5.7950	0.3970	6.9940	1.2710	1648	0.2130	0.4710	8.8890	6.5260
2	10.3960	0	6.7500	8.6690	0.3870	10.3960	1.1970	1347	0.1640	0.4120	12.4400	9.2500
3	10.1480	0	6.6830	8.5360	0.3480	10.1480	1.1310	941.8000	0.1160	0.3830	11.4760	8.1110
4	9.8980	0	6.6170	8.4020	0.3090	9.8980	1.0600	458	0.0570	0.3470	10.4940	6.9480
5	9.4150	0	6.4880	8.1440	0.2330	9.4150	0.9150	-732	-0.0930	0.2600	8.6170	4.7310
6	9.2660	0	6.4500	8.0650	0.2100	9.2660	0.8670	-1202	-0.1530	0.2260	8.0350	4.0410
7	1.0900	0.0630	0.4410	0.5130	0.2310	1.0270	2.6510	1.3847e+03	0.6230	0.8040	2.7230	2.2490
8	1.5200	0.0510	0.8210	0.8790	0.2340	1.4680	2.1900	1.4535e+03	0.5430	0.7540	3.2160	2.5720

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
1	1.8832e+03	-2.7728e+...	41.5556	-123.5757	-70.1217	-2.6990e+...	-25.7381	-0.3449	-129.1222	1.9086	7.3578e+04	0.4535

Рис. 1. Интерфейс программного комплекса

Как только сеть обучена, она используется для прогноза значения необходимого коэффициента. Результатом выполнения программы является оценка финансового состояния предприятия по основным коэффициентам. Пример оценки финансового состояния в соответствии с рис. 1 в правом нижнем окне.

Данное приложение позволяет загружать данные в виде файлов с расширением *.dip, в которых содержится вектор входа. Обучение сети осуществляется на наборах предварительно кластеризованных данных (см. рис. 2).

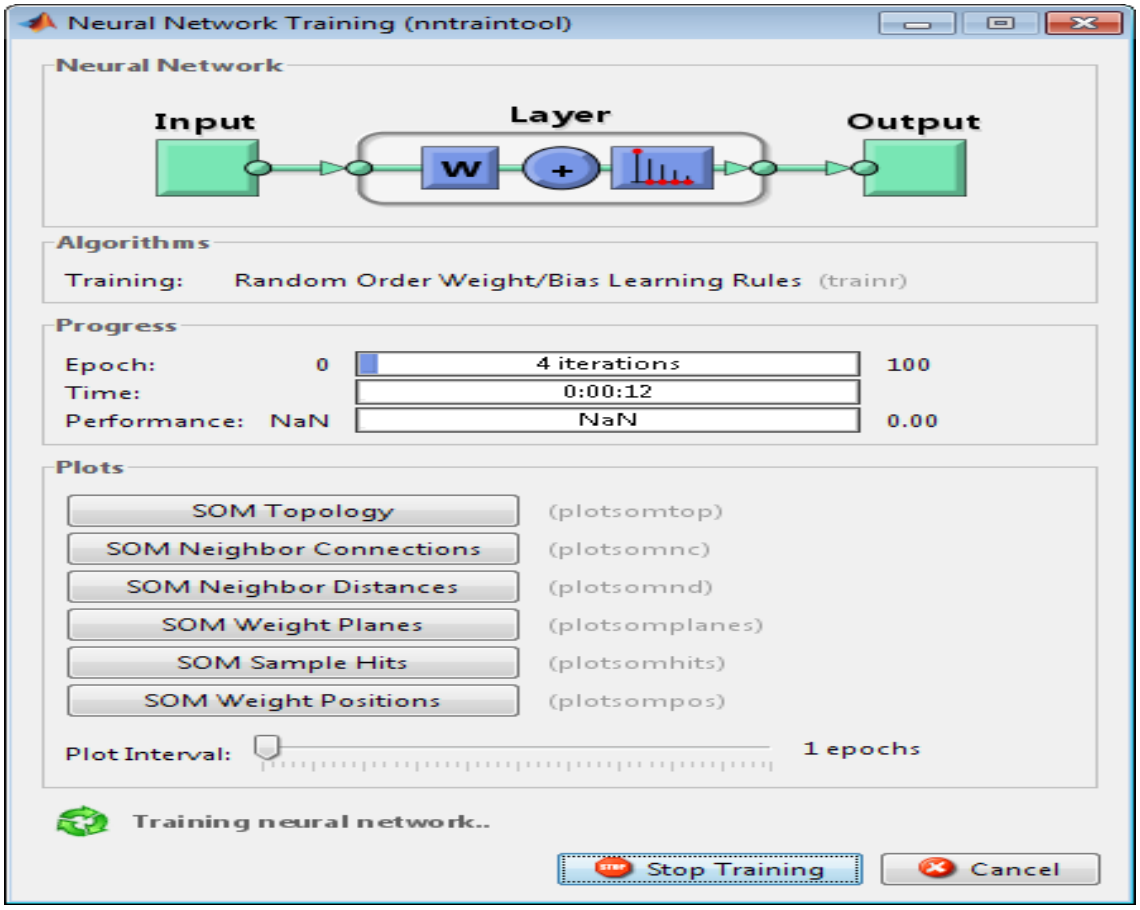


Рис. 2. Процесс обучения нейронной сети

Интерфейс, представленный на рис. 2, создан на основе следующих функций пакета Neural Network Toolbox системы MATLAB: simplefit_dataset, feedforwardnet, train, view, net, perform.

В качестве базовой модели выбрана динамическая искусственная нейронная сеть с прямым распространением сигнала и оценкой качества обучения по среднеквадратической ошибке.

Процесс обучения требует набора примеров ее желаемых выходных данных – входов и желаемых (целевых) выходов.

Динамическая нейронная сеть содержит в своем составе линии задержки, вход нужно рассматривать как последовательность векторов, подаваемых на сеть в определенные моменты времени. Т.е. прогноз зависит не только от одного входного вектора, но и от предыдущего, тем самым реализуется запоминание сети предыдущих данных.

Проверка качества обучения отражает способность сетей выявлять скрытые закономерности в изменении финансовых показателей и, как следствие, способность адекватно прогнозировать значения коэффициентов финансовой устойчивости.

Список использованных источников

1. Мандель И. Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
2. Аббакумов А.А. Разработка методики и алгоритмов идентификации отклонений от нормативов параметров качества электроэнергии в системах электроснабжения: дис. …канд. тех. наук. Саранск, 2005. 180 с.
3. Медведев В. Г. Нейронные Сети Matlab 6. М.: ДИАЛОГ-МИФИ, 2002. 496 с.
4. Калан Р. Основные концепции нейронных сетей: пер. с англ. М.: Вильямс, 2001. 287 с.
5. Афонин В.В., Федосин С.А., Иконников С.Е. Основы теории управления: лабораторный практикум. Саранск: Изд-во Мордов. ун-та, 2008. 244 с.
6. Афонин В.В., Федосин С.А. О структурировании лабораторно-практических занятий при изучении дисциплин программирования // Образовательные технологии и общество. 2014. Т. 17. № 4. С. 497-506. URL:http://cyberleninka.ru/article/n/o-strukturirovanii-laboratorno-prakticheskih-zanyatiy-pri-izuchenii-distsiplin-programmirovaniya (дата обращения 20.10.2015).