

Построение нейросетевых моделей типа «регрессия-авторегрессия» на основе аналитической платформы Deductor

Осипов Геннадий Сергеевич

Сахалинский государственный университет

д.т.н., заведующий кафедрой Информатики

Вашикидзе Нателла Семеновна

Сахалинский государственный университет

доцент кафедры Информатики

Аннотация

Изложены методологические основы построения моделей регрессии и авторегрессии в нейросетевом логическом базисе. Обоснован выбор в качестве инструментария для синтеза модели аналитической платформы *Deductor* компании *BaseGroup Labs*, занимающей лидирующие позиции в отечественном секторе технологий анализа данных. Приведен пример синтеза нейросетевой модели регрессии-авторегрессии для решения задачи прогнозирования.

Ключевые слова: уравнение регрессии, нейронная сеть, прогнозирование.

Construction of neural network models of the «regression-autoregression» type based on the analytical platform Deductor

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

The methodological foundations of constructing regression and autoregressive models in a neural network logical basis are outlined. The selection of the analytical platform *Deductor* of *BaseGroup Labs*, which occupies a leading position in the domestic sector of data analysis technologies, is chosen as a tool for synthesis. An example of synthesis of the neural network regression-auto regression model for solving the prediction problem is given.

Keywords уравнение регрессии, нейронная сеть, прогнозирование

Введение

Важнейшей проблемой современного этапа развития Информатики является задача получения новых знаний из массивов данных, хранящих в себе «большие структуры» информации, характеризующие модели поведения сложных систем. Такие системы часто являются трудно формализуемыми, но, тем не менее, информация об их функционировании может быть использована для изучения и прогнозирования поведения на перспективу. По сути в имеющейся «наблюдаемой» информации, по определению, и заложена модель изучаемой системы. Проблема только в том, чтобы раскрыть (развернуть, распаковать) информацию и заставить ее служить прообразом модели поведения системы.

Известно, что основные достижения в области Искусственного интеллекта связаны с развитием его бионического направления, постулирующего, что любое «мыслящее устройство» должно тем или иным образом воспроизводить структуру человеческого мозга. Таким образом развились подходы к реализации систем искусственного интеллекта на базе нейросетевых и нейро-нечетких парадигм.

Поэтому настоящая работа посвящена проблеме обработки больших массивов информации с целью извлечения из них модели регрессионного типа с использованием многослойной нейронной сети.

Постановка задачи

Объектом и предметом исследования является проблема построения функции f регрессии (авторегрессии) вида [1]:

$$f[\mathbf{x}(\tau), y(\tau)]_{\tau=0,t} = y(t+1) | t \geq 0$$

где $\mathbf{x}(\cdot)$ – вектор внутренних аргументов;

$y(\cdot)$ – прогнозируемая величина.

Целью – отработка методологии синтеза искомой функции f методом извлечения новых знаний из базы данных с помощью нейронной сети [2].

Инструментарий

Проблема синтеза регрессионных моделей достаточно хорошо изучена и формализована. Существуют мощные программные средства, позволяющие решать данную задачу, но в «классической» алгебраической постановке.

Предметная область «Технология анализа данных» развивается благодаря теоретическим исследованиям, находящим практическое воплощение в уникальных специализированных программах.

Аналитической платформы *Deductror* [3] представляет собой многофункциональную комплексную универсальную исследовательскую и прикладную систему, интегрирующую в себе реализацию практически всех существующих современных технологий анализа данных:

Data Mining;

Big Data;

Data Warehousing;

OLAP;
 Business Intelligence;
 Scoring;
 Forecasting;
 Data Quality;
 Machine Learning.

Благодаря доступности и отличному сопровождению данный программный комплекс выбран в качестве инструментария для выполняемого исследования.

Практическая реализация

Решим задачу, сформулированную в [1] – получим прогнозную модель, позволяющую предсказать значение курса акций эмитента на основании обучающей выборки массива исходных данных из предыстории. По сути решается задача восстановления модели (получения нового знания) на основании информации, генерируемой изучаемой системой.

На рисунке 1 представлены начальный и конечный фрагменты используемых данных.

Date	Open	High	Low	Amount	Volume	Close	Close+
22.07.2013	243,99	244,74	241,06	2541960	617617305,1	243,25	243,73
23.07.2013	244,8	246,9	243,5	2966460	727341922,9	243,73	242,65
24.07.2013	243,52	244,72	242,11	2712340	659072971,4	242,65	240,82
25.07.2013	242	242,64	240,03	2103330	507203050,3	240,82	240,87
26.07.2013	241,75	243,5	239,15	2594210	626355852,6	240,87	239,1
29.07.2013	241,18	242,28	238,38	2936390	703924933,3	239,1	236,1
14.07.2017	318	319,55	316,6	3446200	1097277610	318	315,9
17.07.2017	319	319,85	310,75	6157270	1939588227	315,9	315,1
18.07.2017	315,15	317,7	312,75	4402740	1388166557	315,1	318,5
19.07.2017	314,75	318,5	313,8	2654440	839845251	318,5	317,85

Рис. 1 Фрагменты обучающей выборки

В данном случае $x = (Open, High, Low, Close, Amount, Volume)$,
 $y = (Close, Close +)$ – текущая и прогнозная цена закрытия.

На рисунке 2 представлена статистика по обучающей выборке, а на рисунке 3 «расшифровка» гистограммы распределения для (например) параметра *Close*.

Метка столбца	Статистика: Кол-во значений = 1003				
	Гистограмма	Минимум	Максимум	Среднее	Стандартное откл.
9.0 Close		193,2	419,5	275,5822632	49,82179783

Рис. 2 Показатели статистики по исходным данным

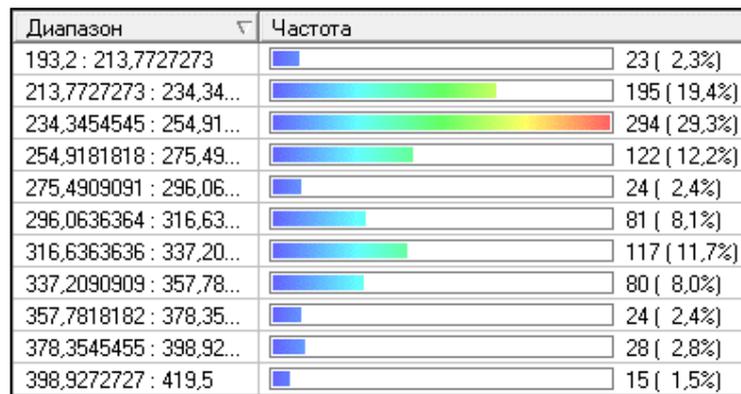


Рис. 3 Подробные данные по гистограмме

На рисунке 4 представлены результаты анализа качества данных.

№	Столбец	Тип данных	Вид данных	Пропуски		Выбросы		Экстремальные		Кол-во уника...	Качество данных	Резюме
				Кол-во	Действие	Кол-во	Действие	Кол-во	Действие			
1	Open	9.0	Веществ...	—	Непрер...						0,8395	Пригоден
2	High	9.0	Веществ...	—	Непрер...						0,8356	Пригоден
3	Low	9.0	Веществ...	—	Непрер...						0,8322	Пригоден
4	Amount	9.0	Веществ...	—	Непрер...		12	Ограничи...	3	Ограничи...	0,5906	Предобработка
✓ 5	Volume	9.0	Веществ...	—	Непрер...		10	Ограничи...	2	Ограничи...	0,4834	Предобработка
6	Close	9.0	Веществ...	—	Непрер...						0,8415	Пригоден
7	Close+	9.0	Веществ...	—	Непрер...						0,8419	Пригоден

Рис. 4 Оценка качества данных

В данном случае существуют выбросы по двум параметрам (*Amount* и *Volume*), которые рекомендуется подвергнуть предобработке. Например, на рисунке 5 представлена гистограмма распределения и график величины *Volume*.

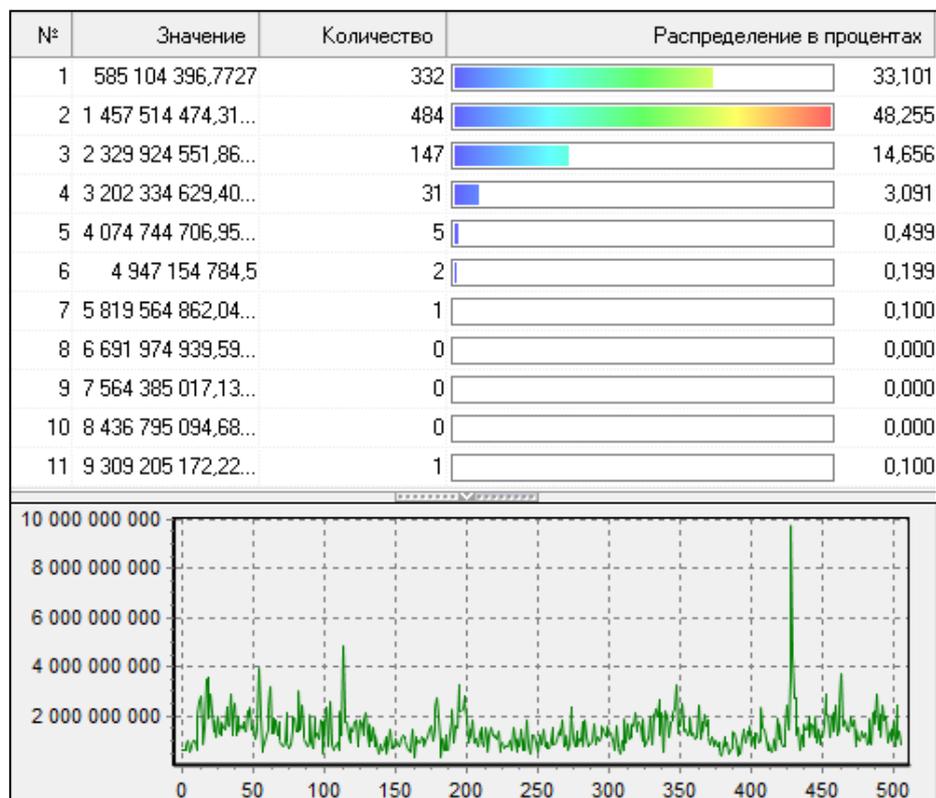


Рис. 5 Подробная расшифровка по величине *Volume*

Возможности аналитической платформы позволяют проводить многогранный анализ используемых данных, например, производить отбраковку незначущих факторов. Так на рисунке 6 приведен отчет по корреляционному анализу.

Входные поля		Корреляция с выходными полями	
№	Поле	Close+	Close-
1	Open	0,992	
2	High	0,994	
3	Low	0,994	
4	Amount	-0,159	
5	Volume	0,174	
6	Close	0,996	

Рис. 6 Расчет корреляций

Далее осуществляется настройку сети – задание топологии (число скрытых слоев и нейронов в ней), выбор функции активации, фиксация метода обучения, разбиение исходного массива данных на обучающее и тестовое множество, формулирование условий останковки ее обучения.

На рисунке 7 представлен граф, определяющий топологию построенной и обученной сети, на котором цвета линий характеризуют веса синаптических связей между нейронами.

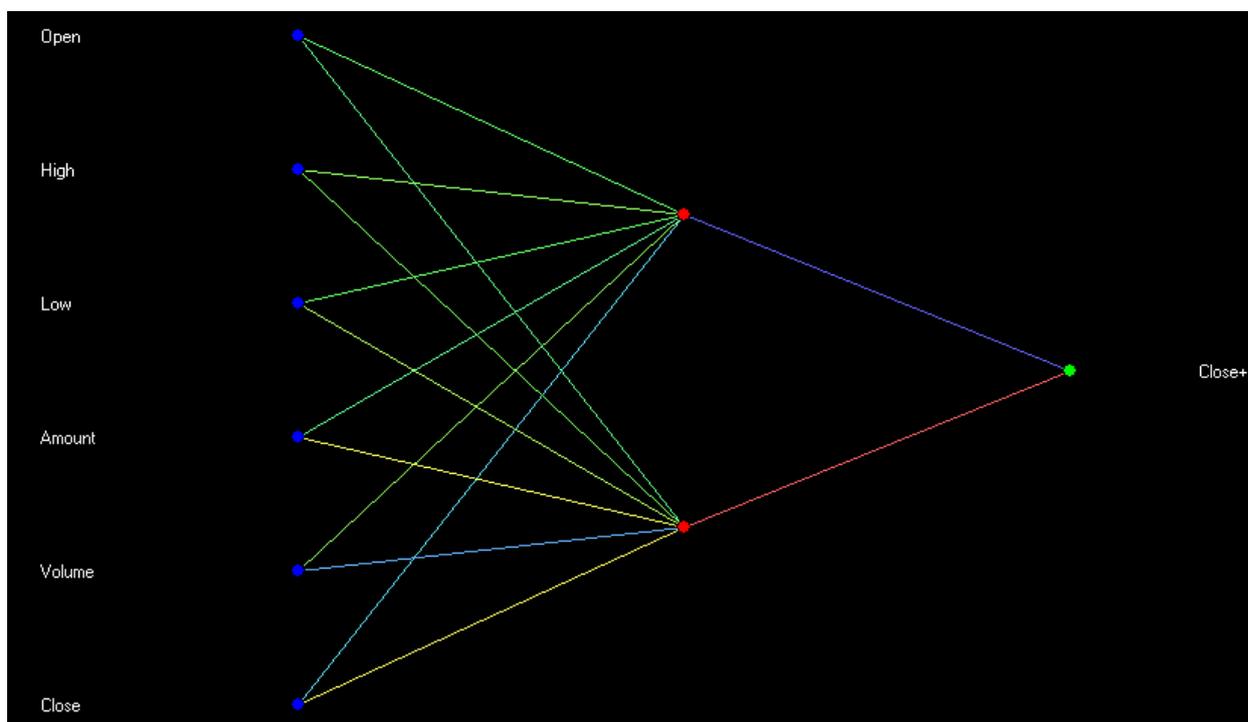


Рис. 7 Топология сети

Качество обучения может характеризоваться диаграммой рассеяния, представленной на рисунке 8.

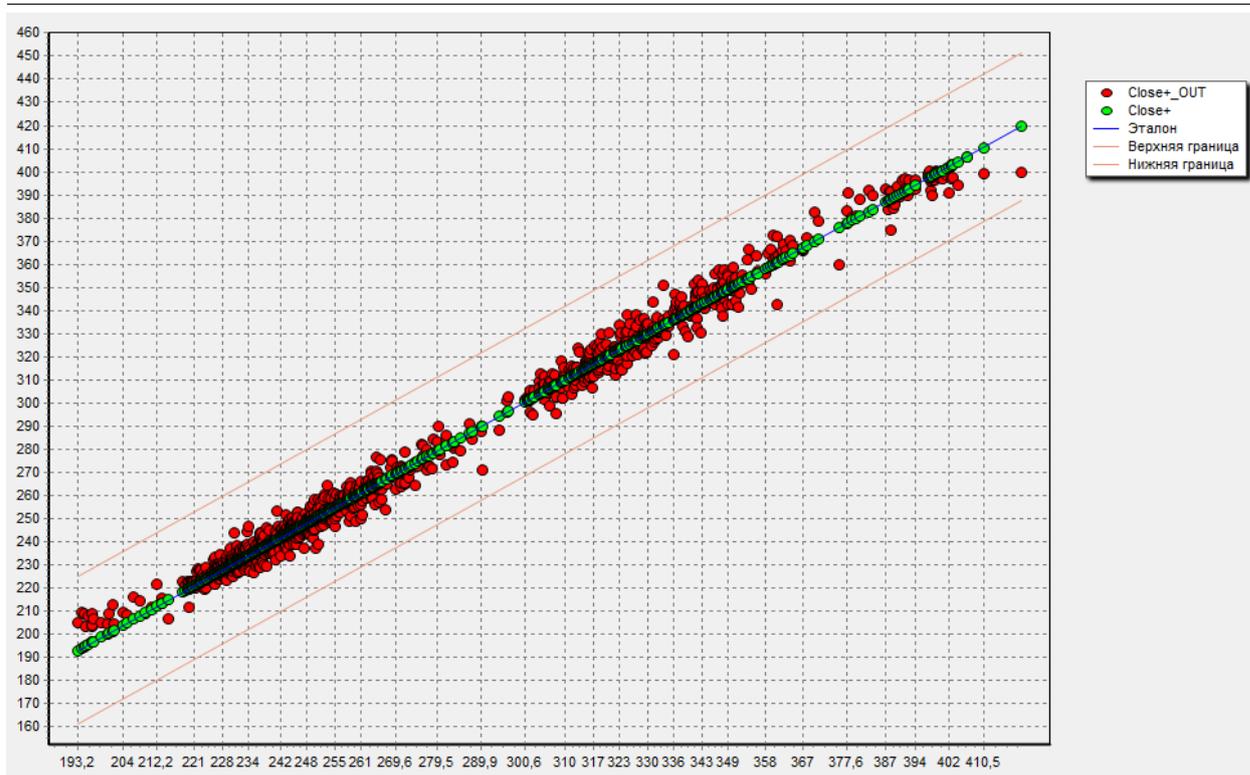


Рис. 8 Диаграмма рассеяния

Обученная сеть может использовать, например, для прогнозирования цены закрытия на следующий рыночный день. Так на рисунке 9 представлен фрагмент диаграммы «Что-если», где предсказано значение цены закрытия на следующий день (на 20.07.2017). Известно, что реальное значение прогнозируемой величина на данный день составило 317,4.

Поле	Значение
Входные	
9.0 Open	319
9.0 High	322,5
9.0 Low	317
9.0 Amount	3491430
9.0 Volume	1117478344
9.0 Close	317,85
Выходные	
9.0 Close+	317,636329221919

Рис. 9 Результат прогноза

Выводы

Проведенное исследование позволяет сделать следующие выводы:

1. Извлечение знаний из данных эффективно решается с помощью искусственных нейронных сетей, формально-топологически идентичных строению человеческого мозга.

2. Решение задач прогнозирования, основанные на использовании нейросетевых моделей типа регрессия-авторегрессия подтверждает

правильность гипотезы о том, что обученная нейронная сеть является моделью исследуемой системы с достаточной степенью адекватности.

3. Возможности аналитической платформы *Deductor* позволяют в единой оболочке эффективно решать комплекс задач, связанных с технологией обработки данных и извлечения знаний (без программирования).

Библиографический список

1. Осипов Г.С., Вашакидзе Н.С., Филиппова Г.В. Основы прогнозирования финансовых временных рядов на базе NeuroXL Predictor // Постулат. 2017. № 7. С. 26.
2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. 2-е изд. СПб.: Питер, 2013. 704 с.
3. Deductor. Продвинутая аналитика без программирования. URL: <https://basegroup.ru/deductor/description> (дата обращения: 08.08.2017).