УДК 004.89

Исследование влияния соотношения тренировочных и тестовых данных на точность прогнозирования с помощью многослойной нейронной сети

Ким Наталья Георгиевна
Сахалинский государственный университет
Студент

Осипов Геннадий Сергеевич
Сахалинский государственный университет
д.т.н., заведующий кафедрой Информатики

Аннотация
Приведена постановка задачи прогнозирования на больших данных применительно к рынку ценных бумаг. Синтезирована обучающая выборка в виде массива исходной информации о котировках и других параметрах, характеризующих функционирование системы прогнозирования. Разработано программное обеспечение решения задачи прогнозирования цены закрытия на следующий торговый период. Дана рекомендация по выбору оптимального соотношения тренировочных и тестовых данных в процессе обучения нейронной сети для минимизации ошибки прогноза.
Ключевые слова: прогнозирование, многослойная нейронная сеть, обучение

Investigation of the influence of the ratio of training and test data on the accuracy of forecasting using a multilayer neural network

Kim Natalia Georgievna
Sakhalin State University
Student

Osipov Gennady Sergeevich
Sakhalin State University
Doctor of technical Sciences, Head of the Department of Computer Science

Abstract
The formulation of the problem of forecasting on big data in relation to the securities market is given. A training sample was synthesized in the form of an array of initial information about quotes and other parameters characterizing the functioning of the forecasting system. The software for solving the problem of forecasting the closing price for the next trading period has been developed. A recommendation is given for choosing the optimal ratio of training and test data in the process of training a neural network to minimize the prediction error.
Keywords forecasting, multilayer neural network, training

Введение

Современное состояние исследований в области анализа больших данных основано на использовании интеллектуальных систем моделирования и прогнозирования построенных на парадигме искусственных нейронных сетей.

Нейросетевая концепция анализа трудно формализуемых задач обеспечивает реализацию симбиоза бионического и нейрофизиологических подходов к решению сложных проблем, характерных для интеллектуальной деятельности человека.

Постановка задачи

Имеется выборка котировок ценной бумаги за определенный период времени. Многомерный временной ряд оформлен в виде большого массива данных, представленных на рисунке 1.

1	Date	Open	High	Low	Close	Amount	Volume	Close+
2	10.01.2019	243,99	244,74	241,06	243,25	2541960	617617305,1	**243,73**
3	11.01.2019	244,8	246,9	243,5	**243,73**	2966460	727341922,9	242,65
4	12.01.2019	243,52	244,72	242,11	242,65	2712340	659072971,4	240,82
5	13.01.2019	242	242,64	240,03	240,82	2103330	507203050,3	240,87
6	14.01.2019	241,75	243,5	239,15	240,87	2594210	626355852,6	239,1
7	15.01.2019	241,18	242,28	238,38	239,1	2936390	703924933,3	236,1
					…			
1001	05.10.2021	318	319,55	316,6	318	3446200	1097277610	315,91
1002	06.10.2021	319	319,85	310,75	315,91	6157270	1939588227	315,1
1003	07.10.2021	315,15	317,7	312,75	315,1	4402740	1388166557	318,5

Рисунок 1 Фрагмент исходных данных

Здесь x= {Date, Open, High, Low, Close, Amount, Volume} – входные данные,

где:

Open – цена открытия торговой сессии (дня);

High – наивысшая цена за торговый день;

Low – наименьшая цена;

Close – цена закрытия сессии;

Amount – количество сделок (покупок/продаж) с ценной бумагой;

Volume – суммарная стоимость сделок.

y= {Close+} – цена закрытия на следующий день.

Требуется на основании данных x об итогах текущей сессии сделать прогноз цены закрытия y ценной бумаги на следующий день, т.е. установить соответствие $x \rightarrow y$ (или, иными словами, идентифицировать многомерную функцию $f: x \rightarrow y$).

Таким образом целью настоящего исследования является разработка программной системы, позволяющей решать задачу прогнозирования (предсказания) применительно к рынку ценных бумаг

Метод решения

Для решения задачи прогнозирования на больших данных используется многослойная нейронная сеть [1]. В качестве инструментария выбрана система символьной математики (компьютерной алгебры) Wolfram Mathematica [2]

Основные результаты

На рисунке 2 представлен фрагмент программы формирования нейронной сети с учетом разделения на тренировочные и тестовые данные в среде Wolfram Mathematica.

```
Нейронная сеть с разделением на тренировочные и тестовые данные

trainIds = RandomSample[Range[Length[x]], Round[Length[x] * 0.95]]; (*индексы для обучающей выборки*)
           случайная выб··· диап··· длина        окру··· длина

valIds = Complement[Range[Length[x]], trainIds];(*индексы для тестовой выборки*)
         дополнение      диап··· длина

net2 = NetInitialize @ NetChain[{
       инициализироват··· нейронная сеть
     LinearLayer["Input" → Dimensions[x][[2]], "Output" → Dimensions[x][[2]],
     линейный слой       размеры массива       размеры массива
      "Weights" → DiagonalMatrix[1 / ArrayReduce[Max, x, 1]], LearningRateMultipliers → None],
                  диагональная матрица  сокращение ···  максимум  множители оценок обучения  ни одного
     LinearLayer["Input" → Dimensions[x][[2]], "Output" → 2],
     линейный слой       размеры массива
     LogisticSigmoid,
     логистическая функция
     LinearLayer["Input" → 2, "Output" → 1],
     линейный слой
     LogisticSigmoid,
     логистическая функция
     LinearLayer["Input" → 1, "Output" → "Real", "Weights" → Max[y]]
     линейный слой                                 максимум
    }];
```

Рисунок 2 Основные определения многослойной нейронной сети

Структура синтезируемой нейронной сети представлена на рисунке 3.

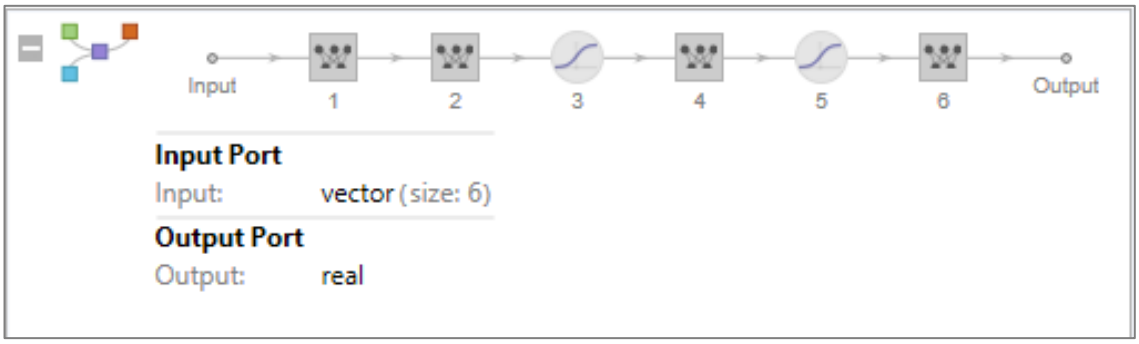

Рисунок 3 Модель сети в среде программирования

На рисунке 4 показаны операторы, обеспечивающие обучение нейронной сети на множестве данных, фрагмент которых представлен на рисунке 1.

```
trainedNet2All = NetTrain[net2, x[[trainIds]] → y[[trainIds]], All, ValidationSet → (x[[valIds]] → y[[valIds]]),
                 тренировать нейронную сеть                    всё  проверочное множество
  MaxTrainingRounds → 4000, TrainingProgressMeasurements → "StandardDeviation",
  максимальное количество рау···  измерения прогресса обучения
  TrainingStoppingCriterion → <|"Criterion" → "Loss", "Patience" → 20|>, BatchSize → 16]
  критерий остановки обучения                                          размер группы данных
```

Рисунок 4 Операторы обучения сети

Процесс обучения с разделением на тренировочные и тестовые множества иллюстрирует рисунок 5

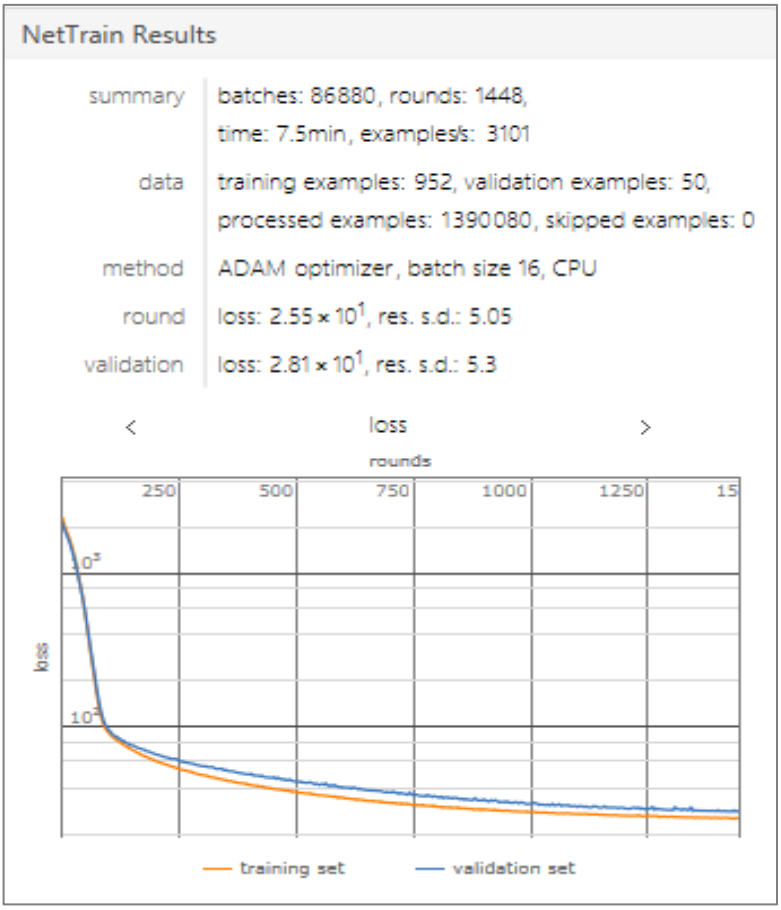

Рисунок 5 Представление процесса обучения сети

Отклонения промежуточных расчетных величин прогноза от ежедневных табличных данных (в процентном выражении по оси ординат) представлены на рисунке 6.

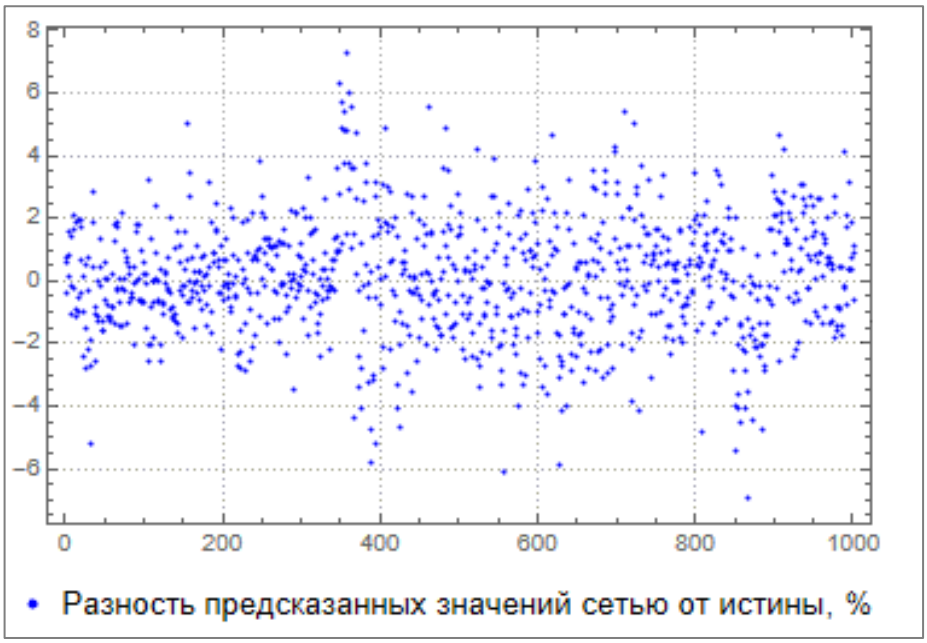

Рисунок 6 Отклонения расчетных величин прогноза от эталона

Соответствующая гистограмма с нанесенной кривой нормального распределения представлена на рисунке 7.

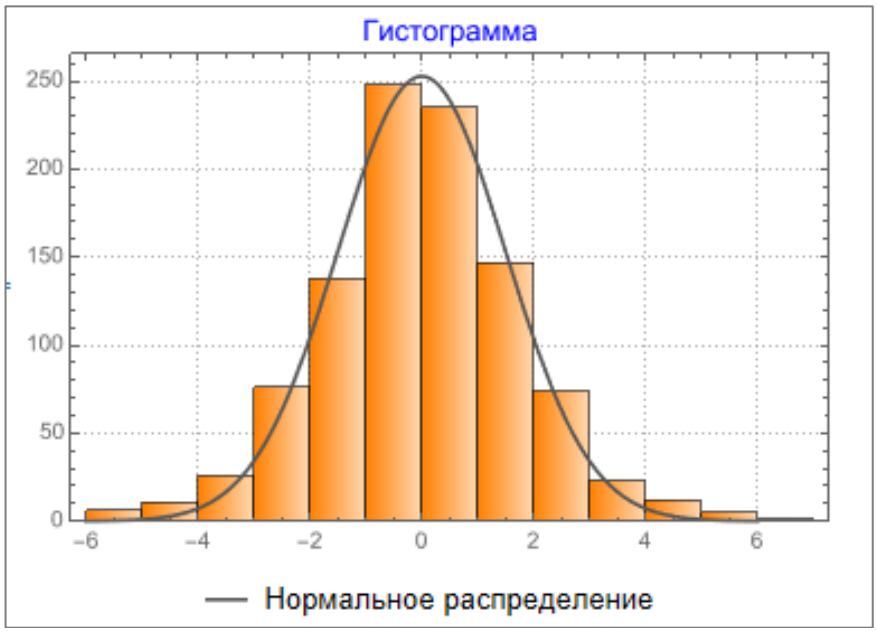

Рисунок 7 Гистограмма и соответствующее нормального распределения

Здесь по оси абсцисс отложены отклонения от эталона в процентном выражении, а по оси ординат – количество величин с таким отклонением.

Окончательные итоги результатов оценки влияния соотношения тренировочных и тестовых данных (при обучении сети) на точность прогноза цены закрытия на следующий день представлены в таблице 1.

Таблица 1– Сравнение результатов прогноза

Показатель	Доля тренировочных данных%				
	80%	85%	90%	95%	100%
Относительная ошибка прогноза	0.41%	0.09%	**0.03**%	0.37%	0.44%
Коэффициент корреляции	99.45	99.45	99.44	99.49	99.16
Среднеквадратичное отклонение	5.25	5.22	5.28	5.03	6.50

Выводы

Проведенное исследование и компьютерное моделирование позволяет сделать следующие выводы:

1. Наименьшая относительная ошибка прогноза (предсказания) цены закрытия на следующий период времени достигается при доле тренировочных данных равной 90%. В этом случае относительная ошибка составляет 0.03%.

2. Программная реализация задачи является унифицированной и может быть использована для решения широкого круга проблем прогнозирования на больших данных в различных предметных областях.

Библиографический список

1. Ким Н.Г., Хлебородова Л.Д. Прогнозирование котировок ценных бумаг методами линейной регрессии, дерева решений и с помощью многослойной нейронной сети // В сборнике: «СТУДЕНТ ГОДА 2021». Сборник статей Международного учебно-исследовательского конкурса. (19 мая 2021 г.) ч.1. Петрозаводск, МЦНП «Новая наука» 2021. С. 288-292. DOI: 10.46916/02062021-4-978-5-00174-249-4
2. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: https://www.wolfram.com/language/elementary-introduction/2nd-ed/ (Дата обращения 07.10.2021).