

УДК 004.85

ПОНИЖЕНИЕ РАЗМЕРНОСТИ ОБУЧАЮЩЕЙ ВЫБОРКИ ПРИ РЕШЕНИИ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ С ПОМОЩЬЮ МНОГОСЛОЙНОЙ НЕЙРОННОЙ СЕТИ

Ким Наталья Георгиевна
студент

Научный руководитель: **Осипов Геннадий Сергеевич**
д.т.н., зав. кафедрой Информатики
ФГБОУ ВО «Сахалинский государственный университет»

Аннотация: Проведено исследование проблемы понижения размерности обучающей выборки без снижения ее информативности. Выполнено компьютерное моделирование и практическая апробация методологии понижения размерности задачи на базе аналитической платформы символьной математики. Проведен анализ влияния на процесс обучения нейросети разделения обучающей выборки на обучающее и тестовое множества.

Ключевые слова: прогнозирование на больших данных, многослойная нейронная сеть, обучение.

REDUCING THE DIMENSION OF THE TRAINING SAMPLE WHEN SOLVING THE FORECASTING PROBLEM USING A MULTILAYER NEURAL NETWORK

Kim Natalia Georgievna
Scientific adviser: Osipov Gennady Sergeevich

Abstract: The study of the problem of reducing the dimension of the training sample without reducing its informativeness. Computer modeling and practical approbation of the methodology for reducing the dimension of the problem on the basis of the analytical platform of symbolic mathematics are carried out. The analysis of the impact on the learning process of the neural network of the separation of the training sample into training and test data is carried out.

Key words: big data forecasting, multi-layer neural network, training.

Объект исследования

Объектом и предметом исследования является проблема снижения размерности задачи прогнозирования с помощью многослойной нейронной сети с целью минимизации трудоемкости поиска решения без потери информативности используемой обучающей выборки.

Целью – отработка методологии понижения размерности в системе компьютерной алгебры Wolfram Mathematica [1], которая является одной из наиболее современных унифицированных платформ компьютерного моделирования, решения задач машинного обучения и разработок в области синтеза систем искусственного интеллекта на базе нейронных сетей.

Обучающая выборка

Задача прогнозирования решается на базе обучающей выборки (за определенный период торговых сессий) котировок ценной бумаги, оформленной в виде большого (более 1000 наблюдений) массива исходных данных [2, 3].

Введем следующие обозначения:

1. k – размерность задачи – число переменных;
2. m – число нейронов в скрытом слое;
3. T – продолжительность обучения сети.

Методология и методы решения

Задача прогнозирования решалась с помощью искусственной многослойной нейронной сети, в которой k нейронов во входном слое, m нейронов в скрытом слое и один нейрон в выходном слое – результат прогноза на следующий период.

Для контроля качества предсказания использовалась ошибка экстраполирования (Relative Extrapolation Error):

$$REE = \frac{|y - y^*|}{y^*} \cdot 100\%,$$

где y – значение цены закрытия на текущий (предшествующий прогнозируемому) день, y^* – эталонное (проверочное) значение цены.

На рисунке 1 представлен фрагмент текста по созданию в среде Wolfram Mathematica многослойной нейронной сети прямого распространения, например, с $m=5$ нейронами в скрытом слое. На входе в сеть используется функция нормирования исходных данных «на единицу», а на выходе – преобразование результата прогноза в реальную величину.

```

net5 = NetInitialize @ NetChain[{
  |инициализировать... |нейронная сеть
  LinearLayer["Input" → Dimensions[xdef][2], "Output" → Dimensions[xdef][2],
  |линейный слой |размеры массива |размеры массива
    "Weights" → DiagonalMatrix[1 / ArrayReduce[Max, xdef, 1]], LearningRateMultipliers → None],
  |диагональная матрица |сокращение ... |максимум |множители оценок обучения |ни одного
  LinearLayer["Input" → Dimensions[xdef][2], "Output" → 5],
  |линейный слой |размеры массива
  LogisticSigmoid,
  |логистическая функция
  LinearLayer["Input" → 5, "Output" → 1],
  |линейный слой
  LogisticSigmoid,
  |логистическая функция
  LinearLayer["Input" → 1, "Output" → "Real", "Weights" → Max[y]]
  |линейный слой |максимум
}]; (*Первый слой – нормирование данных, последний – приведение к реальным масштабам*)

```

Рис. 1. Синтез нейронной сети

Основные результаты

А. Обучение без разделения обучающей выборки на обучающее и тестовое множества.

На рисунке 2 представлены основные операторы обучения нейронной сети, которая аппроксимирует отображение $x \rightarrow y$ (вход $\rightarrow$ выход).

```

trainedNet5 = NetTrain[net5, xdef → y, All, LossFunction → MeanSquaredLossLayer[],
  |тренировать нейронную сеть |всё |функция потерь |слой для расчёта средней абсо
  TrainingProgressMeasurements → "StandardDeviation", BatchSize → 16]
  |измерения прогресса обучения |размер группы данных

```

Рис. 2. Организация обучения нейронной сети

Показатели критерия качества обучения REE и время T, затрачиваемое на обучение при $m=2, 5, 7$ и $k=6, 4, 2$ приведены на рисунке 3.

REE	k	6	4	2		T	k	6	4	2
m	2	1,20%	0,68%	0,15%		m	2	40,6	43,7	44,3
	5	0,92%	0,50%	0,21%			5	41,8	45,4	43,3
	7	0,90%	0,45%	0,19%			7	42,9	44,3	42,6

Рис. 3. Оценка качества обучения

Оба показателя качества решения приводятся к нечетким соответствиям и нормализуются (см. рисунок 4).

1,00	0,50	0,00		0,00	0,65	0,77
0,74	0,34	0,06		0,24	1,00	0,56
0,71	0,28	0,04		0,49	0,76	0,42

Рис. 4 Нечеткое представление показателей обучения

Далее, используя стандартную треугольную конорму $S(a,b) = a + b - ab$, синтезируем интегральный показатель качества и выбираем наилучшее решение. На рисунке 5 справа представлено итоговое нечеткое соответствие с нормализованными значениями функций принадлежности.

1,00	0,83	0,77		1,00	0,69	0,58
0,80	1,00	0,59		0,64	1,00	0,25
0,85	0,83	0,45		0,73	0,69	0,00

Рис. 5 Итоговое нечеткое соответствие

В данном случае наилучшим решением является использование $m = 7$ нейронов в скрытом слое и снижение размерности задачи до $k = 2$.

В. Обучение с разделением обучающей выборки на обучающее и тестовое множества.

Операторы разделения обучающей выборки на обучающее и тестовое множества приведены на рисунке 6.

```

Нейронная сеть с разделением на тренировочные и тестовые данные;

trainIds = RandomSample[Range[Length[xdef]], Round[Length[xdef] * 0.9]];
           |случайная выб... |диап... |длина           |окру... |длина
           (*индексы для обучающей выборки*)

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

```

Рис. 6 Разделение обучающей выборки

Для этого случая процесс обучения сети с критерием остановки представлен на рисунке 7.

```

trainedNet5 = NetTrain[net5, xdef[[trainIds]] → y[[trainIds]], All,
    |тренировать нейронную сеть| |всё|
ValidationSet → (xdef[[valIds]] → y[[valIds]]), LossFunction → MeanSquaredLossLayer[],
    |проверочное множество| |функция потерь| |слой для расчёта средней абсол
MaxTrainingRounds → 4000, TrainingProgressMeasurements → "StandardDeviation",
    |максимальное количество раундов| |измерения прогресса обучения|
TrainingStoppingCriterion → <|"Criterion" → "Loss", "Patience" → 20|>, BatchSize → 16]
    |критерий остановки обучения| |размер группы данных|

```

Рис. 7. Операторы обучения сети

Рисунок 8 демонстрирует заключительный этап обучения нейронной сети с заданными основными параметрами – обучающей и тестовой выборками, количеством эпох обучения.

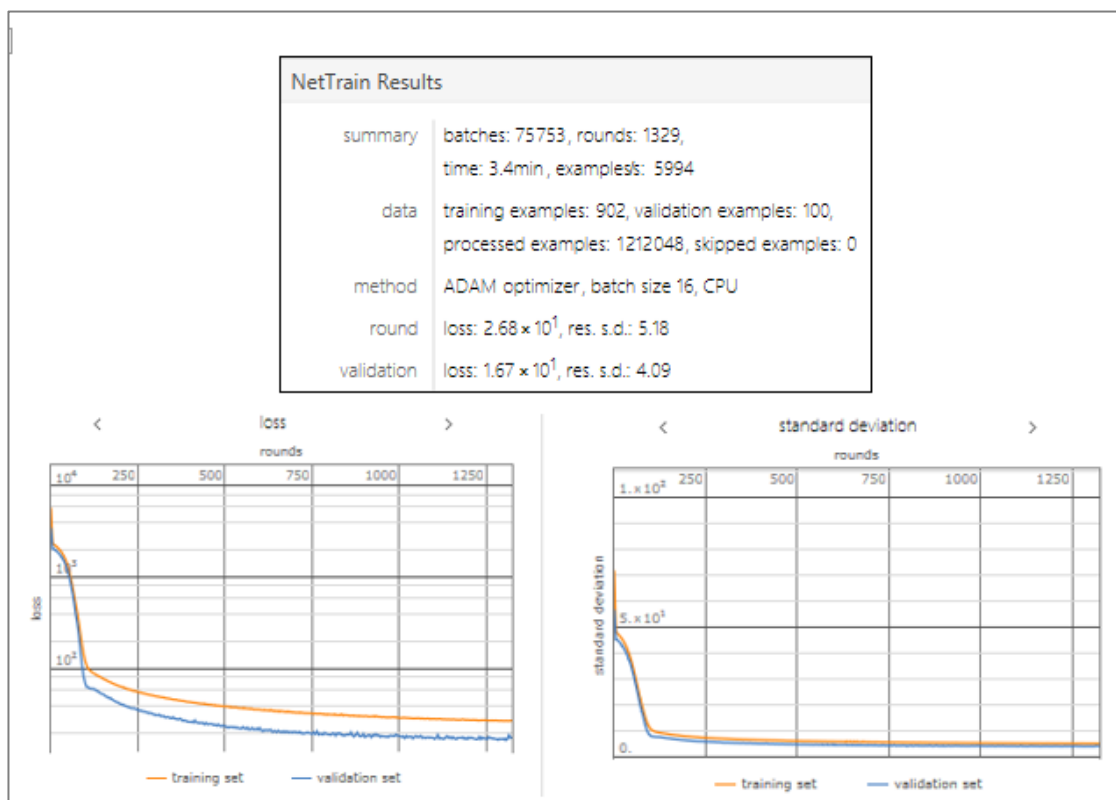


Рис. 8. Иллюстрация процесса обучения сети

Свертка двух показателей качества обучения приведена на рисунке 9 в виде нечеткого соответствия.

0,49	0,99	0,00
0,98	1,00	0,41
0,93	0,98	0,06

Рис. 9. Итоговое нечеткое соответствие

В данном случае (обучение с разделением обучающей выборки на обучающее и тестовое множество) снижение размерности задачи до $k=2$ обеспечивает наилучший вариант решения при $m=2$ нейронам в скрытом слое.

Отметим, что средняя ошибка экстраполирования без разделения обучающего множества на тренировочные и тестовые данные, составляет 0,6%, а с разделением 0,5%.

Выводы

Разработанное программное обеспечение предназначено для исследования проблемы влияния понижения размерности обучающей выборки на трудоемкость решения задачи прогнозирования с помощью многослойной нейронной сети. Кроме того, приведена оценка влияния разделения данных на тренировочные и тестовые.

Показатели качества обучения сведены к одному итоговому критерию, представленному в виде нечеткого соответствия, который позволяет выбрать оптимальный вариант решения задачи.

Список литературы

1. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: <https://www.wolfram.com/language/elementary-introduction/2nd-ed/> (Дата обращения 25.10.2021).
2. Ким Н.Г. Прогнозирование котировок ценных бумаг методами линейной регрессии, дерева решений и с помощью многослойной нейронной сети / Н.Г. Ким, Л.Д. Хлебородова // Студент года 2021: Сборник статей Международного учебно-исследовательского конкурса в 6-ти частях, Петрозаводск, 19 мая 2021 года. – Петрозаводск, 2021. – С. 288-292. – DOI 10.46916/02062021-4-978-5-00174-249-4.
3. Ким Н.Г. Исследование влияния соотношения тренировочных и тестовых данных на точность прогнозирования с помощью многослойной нейронной сети / Н.Г. Ким, Г.С. Осипов // Постулат. № 10 (2021), с. 5.