ПРЕДОБРАБОТКА ДАННЫХ ДЛЯ АНАЛИЗА РЫНКА ЦЕННЫХ БУМАГ С ПРИМЕНЕНИЕМ НЕЙРОННЫХ СЕТЕЙ

Чуклин В.В., группа ИУС-06м
Руководитель доц. Телятников А.О.

Введение

Важным этапом в решении задачи нейросетевого прогнозирования является предобработка обучающей выборки. От ее состава, полноты и качества в значительной мере зависит и качество получаемого прогноза.

Для большинства нейросетей характерно наличие интервала допустимых значений входных сигналов. Функция активации нейронов устанавливает допустимые границы значений исходных данных.

Масштабирование исходных данных в этот диапазон осуществляется при помощи простейшего преобразования – нормализации [1]. Однако при этом не учитываются характеристики закона распределения данных, поэтому при сильной неравномерности закона распределения допустимый диапазон входных сигналов используется очень неравномерно. В нем присутствуют, как слабо заполненные участки, так и участки скученности значений исходной величины [2].

Целью данной статьи является изучение методов повышения качества прогноза за счет повышения равномерности распределения исходных данных.

1. Методы предобработки данных

1.1 Нормировка исходных данных

Процесс отображения всего множества значений исходной величины в заранее заданный интервал называется нормализацией [1]. Данный интервал называется интервалом допустимых значений и определяется функцией активации нейрона. Для различных функций активации эти интервалы будут

различными. Нормировка необходима для эффективного использования интервала максимальной чувствительности функции активации.

Числовые значения сигналов рекомендуется масштабировать и сдвигать так, чтобы весь диапазон значений попадал в интервал допустимых входных сигналов [1]. Это масштабирование задается формулой (1):

$$x' = \frac{(x - x_{\min})(b - a)}{(x_{\max} - x_{\min})} + a, \quad (1)$$

где [a,b] – интервал допустимых значений входных сигналов, $x_{\max}, x_{\min}$ - максимальное и минимальное выборочные значения признака *x*. Предобработку входного сигнала по формуле (1) называют простейшей предобработкой [1].

1.2 Повышение различимости исходных данных

На практике часто встречается ситуация, когда большая часть поступающих на вход нейросети сигналов занимает лишь малую часть диапазона различимых входных сигналов, это приводит к тому, что нейросеть плохо распознает значения входных сигналов и дает низкую точность прогноза. Поэтому при предобработке данных необходимо должное внимание уделять таким аспектам как распределение величин по интервалу значений, с целью повысить различимость данных нейросетью.

В задаче прогнозирования временных рядов с применением нейросетей наиболее часто используются сети с сигмоидальным нелинейным преобразователем, при котором выходное значение нейрона лежит в интервале [0,1].

Интервал допустимых значений входных сигналов [a,b] должен соответствовать интервалу зоны чувствительности функции активации нейрона, поэтому перед повышением равномерности исходных данных необходимо провести их нормализацию по формуле (1). После нормализации исходных данных будем повышать их различимость. Для этого значения необходимо наиболее равномерно перераспределить по интервалу значений

исходных данных. Таким образом, вводится искусственная равномерность распределения данных по интервалу [2].

Преобразование значений исходной величины x_i выполняется в соответствии с плотностью их распределения *p(x)* по диапазону (рис. 1).

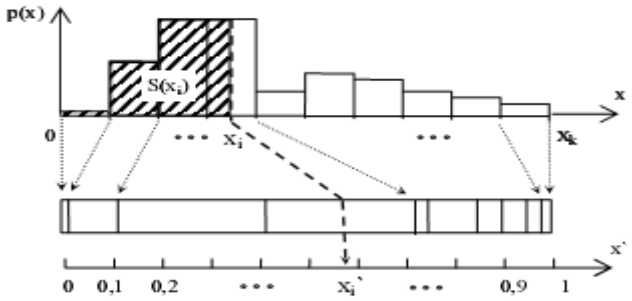


Рисунок 1 – Применение предложенного подхода

Значения преобразованной величины x_i' вычисляется на основании значения исходной величины x_i по формуле (2):

$$x_i' = S(x_i). \tag{2}$$

Физически величине x_i' соответствует площадь $S(x_i)$ фигуры, ограниченная значениями $x_1 = 0$ и $x_2 = x_i$, т.е. с учетом всех предыдущих значений *x*.

$$S(x_i) = P(X < x_i), \tag{3}$$

где *P* – интегральная вероятность значений исходной величины *x*.

1.3 Метод с использованием функции распределения

Данный метод предназначен для повышения различимости данных имеющих определенный закон распределения. Нормальное распределение является одним из наиболее распространенных распределений. Поэтому рассмотрим этот метод на примере нормального закона распределения. Под нормальным распределением подразумевается так называемое стандартное нормальное распределение - нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Произвольную

выборку с другими значениями математического ожидания и дисперсии всегда можно привести к стандартному нормальному распределению.

Функция плотности вероятности стандартного нормального распределения имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right). \quad (4)$$

Интегральная функция вероятности распределения имеет следующий вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2} \right) du\ . \quad (5)$$

Интегральная функция вероятности распределения обычно выражается через специальную функцию *Erf(x)*:

$$Erf(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \exp\left(-u^2\right) du \quad (6)$$

$$F(x) = \frac{1}{2}\left(1 + Erf\left(\frac{x}{\sqrt{2}} \right) \right) \quad (7)$$

Преобразование исходных данных x_i, которые распределены по нормальному закону распределения с параметрами μ и σ, осуществляется следующим образом:

$$x_i' = \frac{1}{2}\left(1 + Erf\left(\frac{x_i - \mu}{\sqrt{2}\sigma} \right) \right) \quad (8)$$

В результате, используя полученное выражение, мы преобразуем исходные данные к равномерному распределению.

2. Экспериментальная проверка методов предобработки

2.1 Объект на котором проверяем методы

Исходными данными служили временные ряды котировок акций компании МТС с 11.02.2004 по 11.09.2007 [3], всего 888 значений по каждой котировке.

В эксперименте использовались приращения значений следующих

котировок: Bid, Ask, цена открытия, минимальная цена, максимальная цена, цена закрытия, средняя цена, оборот за день, объем сделок за день. Гистограммы их плотности распределения приведены на рис 2.

Для каждой переменной были вычислены исходные статистические величины. Данные статистические величины представлены в таблице 1.

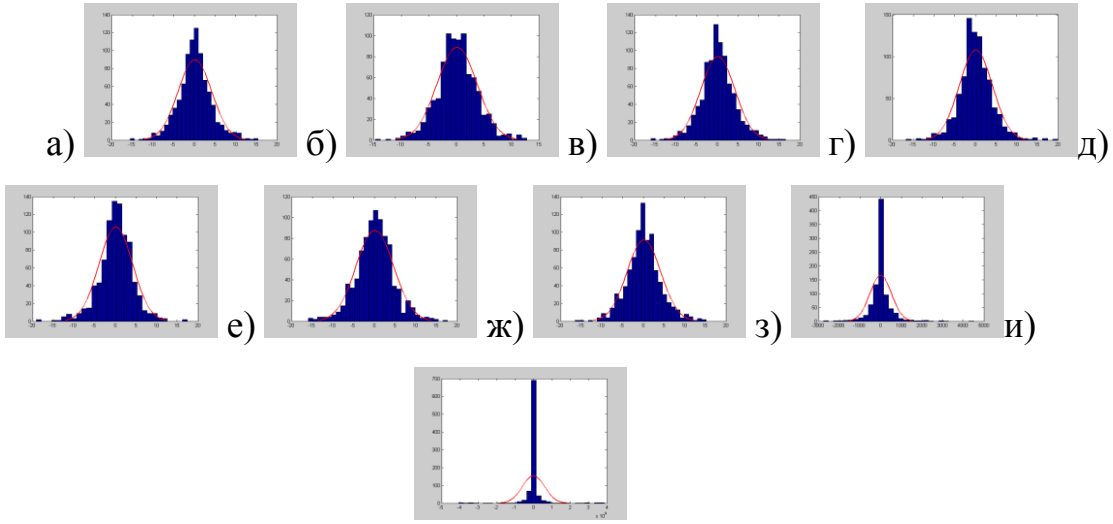


Рисунок 2 – Гистограммы плотности распределения котировок: а) Ask;

б) средняя цена; в) Bid; г) максимальная цена; д) минимальная цена; е) цена открытия; ж) цена закрытия; з) объем сделок за день; и) оборот за день

Таблица 1 – Исходные статистические величины

	M[x]	D[x]	σ[x]	Max	Min
Ask	0,139	16,602	4,075	15,470	-15,510
Средняя цена	0,137	13,311	3,648	12,790	-14,680
Bid	0,140	16,984	4,121	16,5	-16,12
Максим. цена	0,139	16,283	4,035	19,99	-16,79
Миним. цена	0,134	16,614	4,076	17,5	-19,13
Цена открытия	0,134	20,781	4,559	17,77	-16
Цена закрытия	0,139	16,791	4,098	15,22	-16,5
Объем сделок	1,985	2,7031E5	519,910	4510,000	-2777,000

Оборот за день	2,8113E5	3,6505E17	6,0419E8	3,8606E9	-4,0782E9

2.2 Описание нейронной сети

Рассмотренные методы предобработки проверялись на экспериментальных данных. В данном эксперименте прогнозировалось изменение цены закрытия на завтра по пяти предыдущим значениям. Было сформировано 870 примеров для обучающей выборки и 13 примеров для контрольной выборки. Структура нейронной сети: трехслойный персептрон 45-90-1 с сигмоидальной функцией активации, в качестве алгоритма обучения был выбран классический алгоритм обратного распространения ошибки.

2.3 Результаты предобработки исходных данных

Для проведения эксперимента использовались три описанных выше метода: нормировка, повышение различимости, метод с использованием функции распределения вероятностей. Результаты предобработки показаны на гистограммах плотности распределения (рис. 3-5).

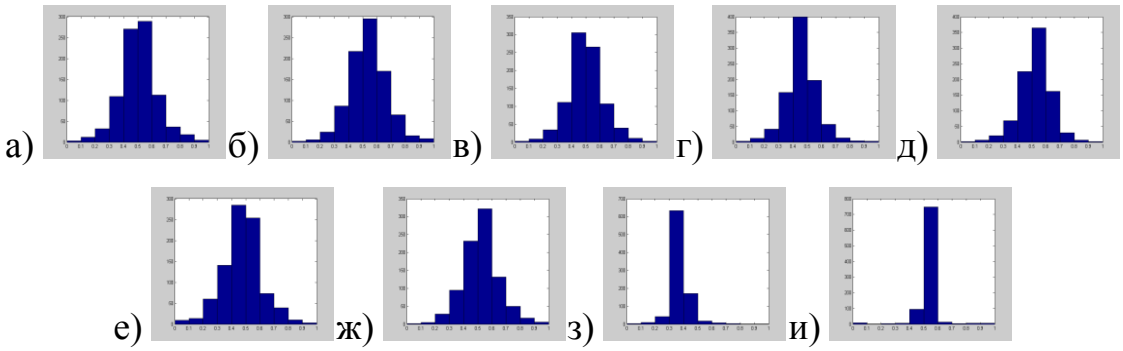


Рисунок 3– Гистограммы распределения котировок, после использования нормировки: а) Ask; б) средняя цена; в) Bid; г) максимальная цена;

д) минимальная цена; е) цена открытия; ж) цена закрытия; з) объем сделок;

и) оборот за день

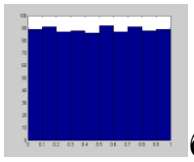
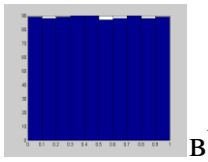
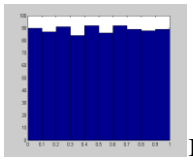
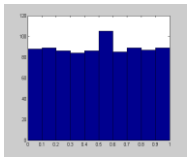
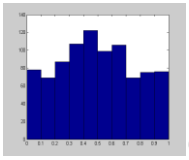
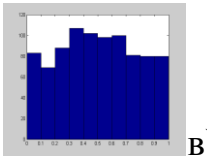
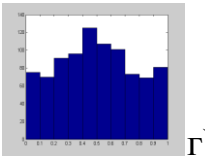
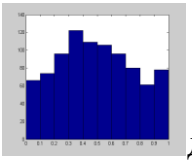
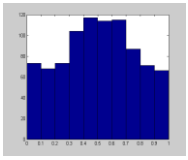
а)  б)  в)  г) 

Рисунок 4 – гистограммы распределения котировок, после использования метода повышения различимости: а) Ask; б) средняя цена; в)объем сделок; г) оборот за день

а)  б)  в)  г)  д) 

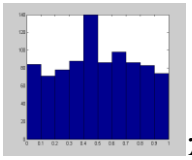
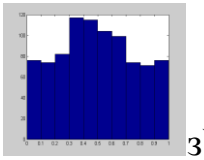
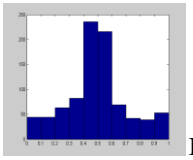
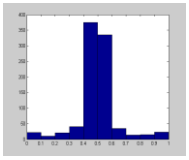
е)  ж)  з)  и) 

Рисунок 5 – Гистограммы распределения котировок, после применения метода предоработки с использованием функции распределения:

а) Ask; б) средняя цена; в) Bid; г) максимальная цена; д) минимальная цена; е) цена открытия; ж) цена закрытия; з) объем сделок; и) оборот за день.

2.4 Результаты на выходе нейронной сети

Обучение нейронной сети проводилось на обучающих выборках предобработанных по трем описанным выше методам, результаты работы сети оценивались на контрольной выборке (таблица 2).

Таблица 2 – Результаты эксперимента

Тип метода предобработки	Средняя ошибка на контрольной выборке
Нормировка	0,47
Повышение различимости	0,26
Применение функции	0,35

распределения	

Наименьшая ошибка прогноза была достигнута на данных предобработанных методом повышения различимости. Полученные результаты приведены на рис. 6.

Выводы

В данной статье для повышения качества прогноза были использованы следующие методы предобработки данных: нормировка, повышение различимости, использование функции распределения.

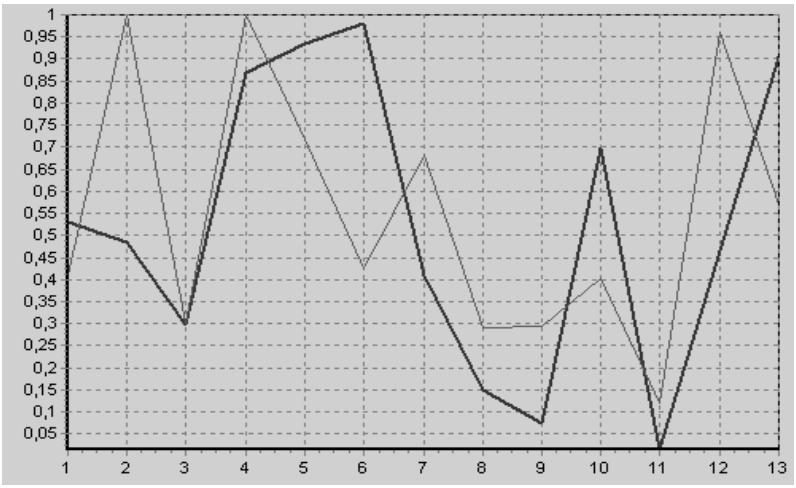


Рисунок 6 – Результат прогнозирования (—— исходные значения; —— прогноз)

Анализируя полученные в результате эксперимента результаты, можно сказать, что применение методов позволяющих повысить равномерность распределения исходных данных по интервалу допустимых значений входных сигналов нейронной сети, улучшает качество прогноза.

Проведенный эксперимент показал, что данные предобработанные методом повышения различимости, наиболее равномерно распределены по допустимому интервалу, при этом на этих данных достигнуто наименьшее значение ошибки прогноза.

Перечень ссылок

1. Миркес Е.М. Нейроинформатика: Учебное пособие для студентов. – Красноярск: ИПЦ КГТУ, 2002. – 347с.

2. Крисилов В.А., Кондратюк А.В. Преобразование входных данных нейросети с целью улучшения их различимости / Электронный ресурс. Способ доступа: URL: http://neuroschool.narod.ru/articles.html

3. Информационный ресурс Investfunds – все о рынке акций / Электронный ресурс. Способ доступа: URL: http://stocks.investfunds.ru/