Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение

высшего образования

«Национальный исследовательский университет

“Высшая школа экономики”»

Жалкова Наталия Евгеньевна

**Задание на прогнозирование**

Отчет студента 2 курса бакалавриата группы №НТ-6

Москва 2017

Работа Жалковой Наталии, группа НТ-6 по теме “Задание на прогнозирование”.

**Задание.**

В репозитории <https://archive.ics.uci.edu/ml/datasets.html> содержится задача «Ирисы Фишера», которая часто используется в качестве эталонной задачи таксономии. С помощью имеющегося в вашем распоряжении нейропакета определите, к какому типу относятся 11 ирисов, характеризующиеся следующими параметрами:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1** | 5,5 | 2,3 | 4,0 | 1,3 | VERSICOL |
| **2** | 6,6 | 3,0 | 4,4 | 1,4 | VERSICOL |
| **3** | 6,8 | 2,8 | 4,8 | 1,4 | VERSICOL |
| **4** | 5,4 | 3,4 | 1,7 | 0,2 | SETOSA |
| **5** | 5,1 | 3,7 | 1,5 | 0,4 | SETOSA |
| **6** | 5,2 | 3,5 | 1,5 | 0,2 | SETOSA |
| **7** | 5,8 | 2,8 | 5,1 | 2,4 | VIRGINIC |
| **8** | 6,7 | 3,0 | 5,0 | 1,7 | VERSICOL |
| **9** | 6,3 | 3,3 | 6,0 | 2,5 | VIRGINIC |
| **10** | 5,3 | 3,7 | 1,5 | 0,2 | SETOSA |
| **11** | 5,0 | 2,3 | 3,3 | 1,0 | VERSICOL |

В крайней правой колонке указан предположительный тип. Проверьте его правильность.

Второй задачей, которую можно решать с помощью того же нейропакета является прогнозирование. Решите обе эти задачи.

**Цель работы:**

Задача 1 - научиться определять тип нового цветка по данным имеющихся измерений.

Задача 2 – решение задач регрессии, в которых выход является непрерывной числовой переменной, с помощью нейронной сети. То есть прогнозировать ширину лепестка.

**Выполнение задания:**

**Задача 1**

1. Преобразую исходные наборы данных в необходимый для вычисления вид, сохранив в .xls и заменив буквенные названия на цифры. Тестовая и обучающая выборки – 7 столбцов(4 для входных данных, 3 – для выходных), в экзаменационной выборке 4 столбика для входных данных. Файлы приложены к отчету.

SETOSA = 1, 0, 0; VIRGINIC = 0, 1, 0; VERSICOL = 0, 0, 1.

SEPALLEN - Длина чашел

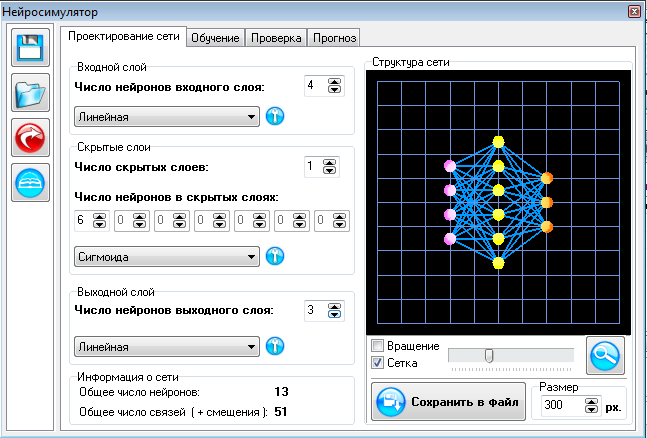
SEPALWID - Ширина чашел

PETALLEN - Длина лепестка

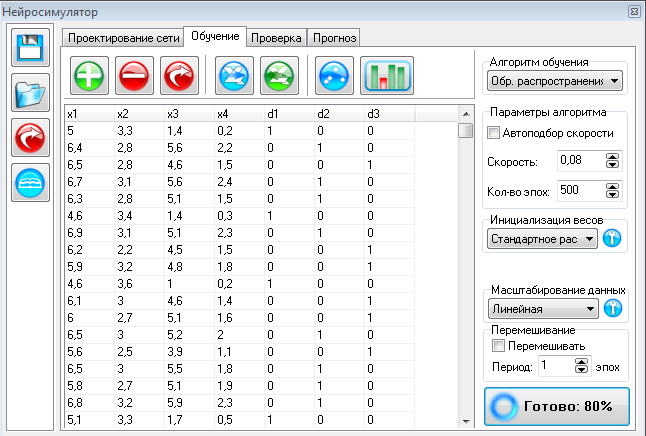
PETALWID - Ширина лепестка

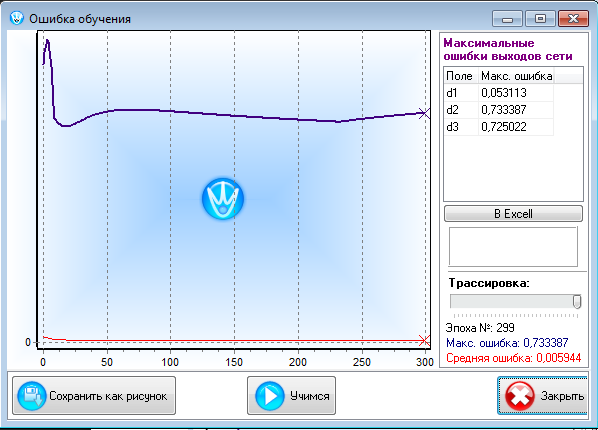
IRISTYPE - Тип ириса

1. Перейдем к созданию нейронной сети. Нейросеть должна иметь 4 входных нейрона и 3 выходных нейрона, 1 скрытый слой с 6 нейронами. Ниже приведен скриншот нейросети, которая показала лучшие результаты.



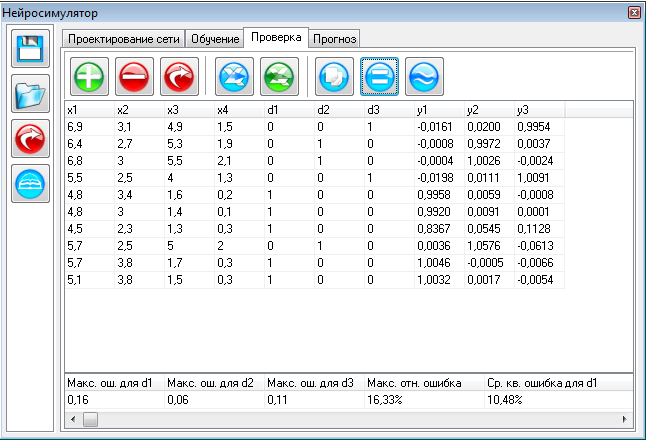
1. Обучим нейросеть, загрузив в соответствующей вкладке подготовленный файл .xls. Получившаяся максимальная ошибка равна 0.7, а средняя 0.006. Так же видно, что наименьшая ошибка получается у сети в расчете d1, то есть SETOSA, что означает то, что точность определения этого вида выше, чем двух других.



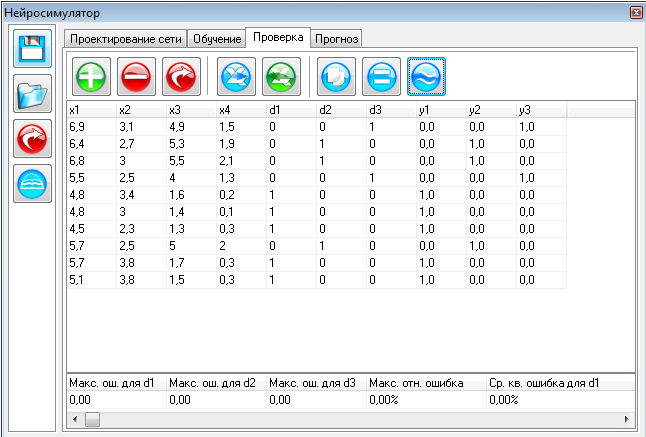


В процессе обучения среднеквадратичная ошибка сильно уменьшается, средняя ошибка остается постоянно на низком уровне, но наибольшая ошибка равна примерно 0.7, что весьма значительно. Это свидетельствует о наличии выбивающихся данных в использующейся выборке.

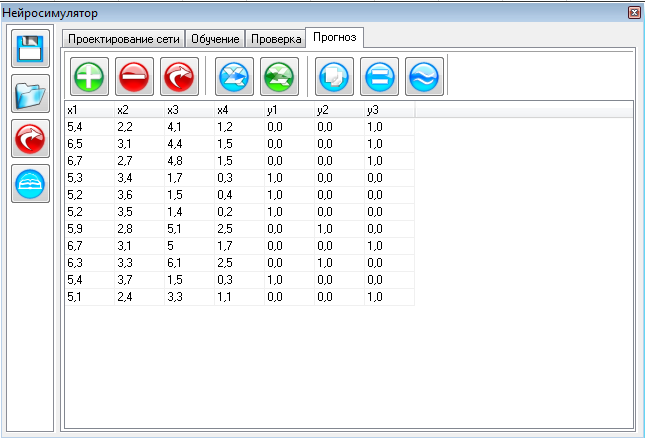
1. Протестирую обученную сеть. Так же, как и в предыдущем пункте, передам в соответствующей вкладке заранее подготовленную таблицу





При проверке же с округлением созданная нейронная сеть выдает идеальные результаты. 

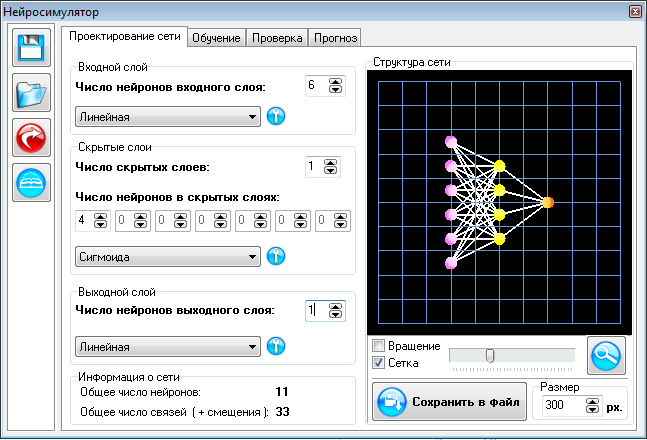
1. Так же проверим экзаменационную выборку. Аналогично двум предыдущим пунктам, во вкладке “Прогноз” загружу подготовленную таблицу



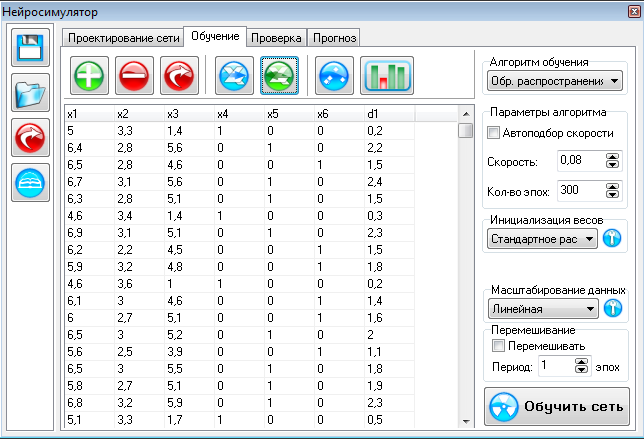
Результаты получились хорошими, если сравнивать данные с данными из обучающей выборки.

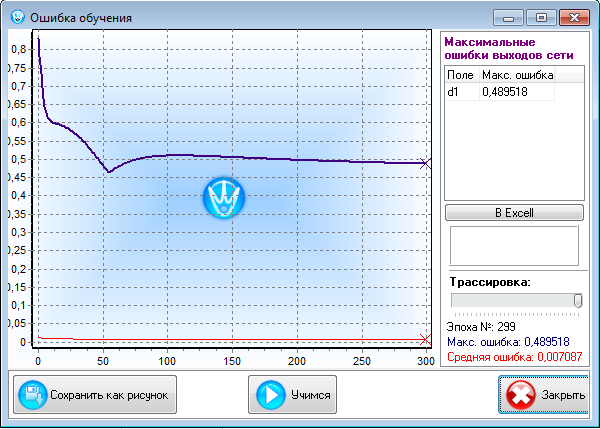
**Задача 2**

1. Исходные выборки возьму из задачи 1, заменив столбцы так, чтобы ширина лепестка стала выходным значением, а тип цветка – входным.
2. Теперь нейросеть должна иметь 6 входных нейрона и 1 выходных нейрона, и 1 скрытый слой с 4 нейронами.



1. Обучение нейросети





Максимальная ошибка довольно таки мала, средняя ошибка – чрезвычайно мала, значит нейросеть успешно обучена.

1. Протестирую нейросеть. Полученные результаты оказались очень близки к правильным. При предсказании ширины лепестка у цветков первого вида среднеквадратичная ошибка очень велика, тогда как у цветков второго и третьего вида предсказания достаточно точны.

Это объясняется разницей в ширине лепестков у первого вида и остальных цветков. Если для него среднее значение ширины лепестка не превосходит 0.3-0.4, то среднее значение ширины лепестка у цветков второго и третьего вида значительно больше.

