Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Исследование и разработка алгоритмов восстановления пропущенных значений в больших массивах данных

Выполнил: Медведев Иван Сергеевич, гр. 7383
Руководитель: Геппенер Владимир Владимирович, д.т.н., профессор
Консультант: Шевская Наталья Владимировна

Санкт-Петербург, 2021

Актуальность

Проблема пропущенных значений достаточно актуальна при проведении экспериментов, например, содержание каких-либо веществ в объектах производства, в социологических опросах, в бухгалтерских отчетах и т.п. Причинами неполноты данных могут быть невнимательность человека, ошибки, поломка оборудования, противоречия результатам экспериментов.

Наличие пропусков может привести к невозможности анализа таких данных или к ошибочным результатам анализа. Поэтому для дальнейшей работы с данными следует заполнять пропуски таким образом, чтобы они не выбивались из общей структуры.

Цели и задачи

Цель: – исследовать алгоритмы нахождения и восстановления пропущенных значений с последующей разработкой собственного алгоритма.

Задачи:

1. провести сравнительный анализ существующих разработок в предметной области;
2. рассмотреть методы решения восстановления пропущенных значений;
3. разработать алгоритм восстановления пропусков в больших массивах данных;
4. оценить абсолютное отклонение, восстановленных при помощи написанного алгоритма, значений
5. определить направления развития.

Сравнительный анализ

Алгоритм	Адаптивность	Не искажает статистические характеристики	Не уменьшает данные
Удаление строк	+	-	-
Замена средним	+	-	+
K-ближайших соседей	+	-	+
ZET-алгоритм	-	+	+
Нейронные сети	+	+	+

Методы решения

Нейронные сети, решающие задачи регрессии:

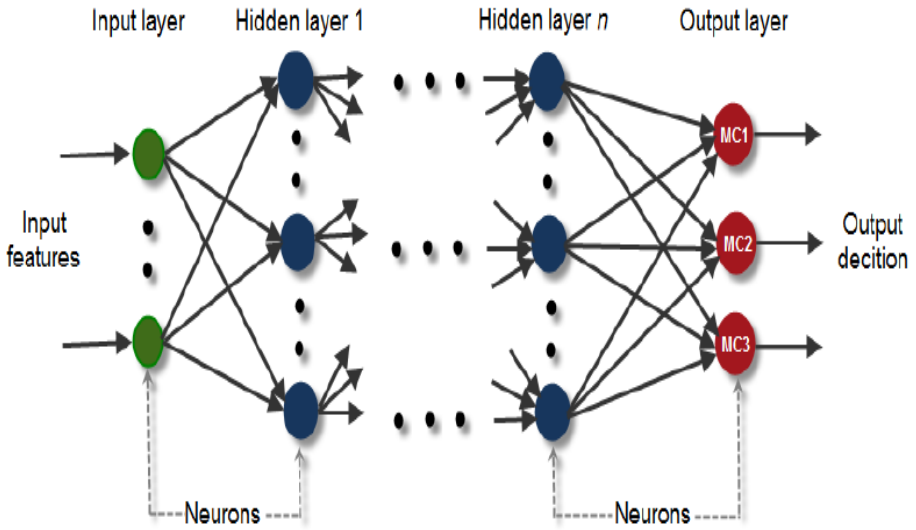

Рисунок 1 – Многослойный перцептрон

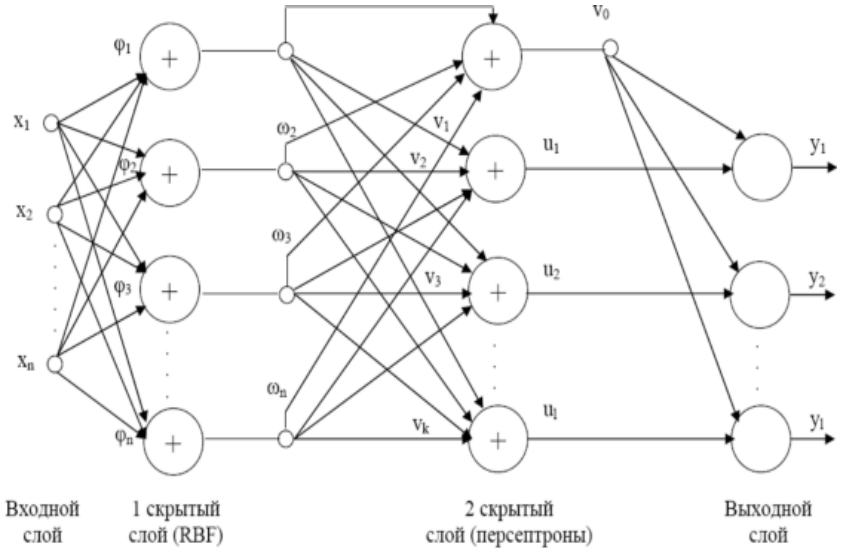

Рисунок 2 – Радиально-базисная нейронная сеть

Обработка данных

Пусть дана матрица *X*, в которой следует восстановить пропуски. Данная матрица разбивается на две подматрицы *X** и *Y*, как показано на рис. 3.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & ? \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & ? & a_{33} & ? \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \Rightarrow \begin{matrix} X^* = \begin{pmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \\ Y = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ? \\ a_{31} & ? & a_{33} & ? \end{pmatrix} \end{matrix}$$

Рисунок 3 – Формирование матриц *X** и *Y*

Далее для каждой строки из матрицы *Y* формируется тренировочный набор, способ формирования представлен на рис. 4.

$$Y_1 = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ? \end{pmatrix} \qquad Y_2 = \begin{pmatrix} a_{31} & ? & a_{33} & ? \end{pmatrix}$$

$$X_1^{train/test} = \begin{pmatrix} a_{21} & a_{22} & a_{23} \\ a_{41} & a_{42} & a_{43} \end{pmatrix} \qquad X_2^{train/test} = \begin{pmatrix} a_{21} & a_{23} \\ a_{41} & a_{43} \end{pmatrix}$$

$$X_1^{labels} = \begin{pmatrix} a_{24} \\ a_{44} \end{pmatrix} \qquad X_2^{labels} = \begin{pmatrix} a_{22} & a_{24} \\ a_{42} & a_{44} \end{pmatrix}$$

Рисунок 3 – Формирование тренировочных наборов

Описание работы алгоритма

Псевдокод работы алгоритма:

Входные данные: двумерный массив с пропусками X

Выходные данные: матрица X с заполненными пропусками

---

normalize(X) // функция, нормализующая столбцы массива.

X^*, Y = splitting_data(X) // функция возвращает подматрицу без строк с пропусками и //подматрицу, состоящую только из строк с пропусками

results = [] //массив с предсказанными значениями

for всех строк Y_i матрицы Y:

X_{train}, X_{test} , L_{train}, L_{test} = get_train_test_data(X^*) // функция, которая возвращает //тренировочные, тестовые наборы и метки

model.fit(X_{train} , $L_{train,,}$ validation_data = (X_{test}, L_{test}))) // обучение сети

results.append(model.predict(Y_i)) // предсказываем значение по строке с пропуском и // добавляем в массив

for каждого пропуска x_i в матрице X:

x_i = results.pop(0) // пропуски заполняем с первого элемента массива results

return X

Абсолютное отклонение нейронных сетей, первый набор данных

Тип / Кол-во Пропусков, %	Перцептрон	Радиально-базисная нейронная сеть	Ансамбль
5	27,239	18,920	21,864
10	38,172	27,453	25,089
20	144,538	110,664	110,543

Абсолютное отклонение нейронных сетей, второй набор данных

Тип / Кол-во Пропусков, %	Перцептрон	Радиально-базисная нейронная сеть	Ансамбль
5	888,631	1466,438	1113,268
10	4023,079	2681,303	3247,584
20	9066,127	7757,36	7440,936

Направление дальнейшего развития

- Модернизировать алгоритм таким образом, чтобы он восстанавливал не только числовые значения, но и категориальные;
- добавить в алгоритм возможность регулирования параметров нейронных сетей;
- Модернизировать алгоритм таким образом, чтобы строки с восстановленными значениями добавлялись в обучающую выборку

Заключение

- Проделанный обзор методов показал, что нейронные сети, по заданным критериям, решают поставленные задачи лучше, чем статистические методы;
- реализован алгоритм для восстановления пропущенных значений, основанный на двух нейронных сетях: перцептрон и радиально-базисная нейронная сеть,
- Экспериментальное исследование показало, что ошибка предсказанных значений не превышает 10% от истинного значения.

Апробация работы

- Репозиторий проекта https://github.com/vanokako/filling_gaps