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

Применение нейросетевого подхода к кластерному анализу распределённых данных

Выполнил: Итальянцев Ян Викторович, гр. 3304
Руководитель: Борисенко Константин Алексеевич, к.т.н., ассистент кафедры МОЭВМ

Актуальность

Работа с Big Data требует применение распределённого подхода к вычислению и работе алгоритмов кластеризации, обученная нейронная сеть позволит автоматизировать этот процесс

Проблематика: кластеризация с использованием данных распределённых между машинами

- требует много вычислительных ресурсов,
- имеет ограничение на количество обрабатываемых данных.

Цель и задачи

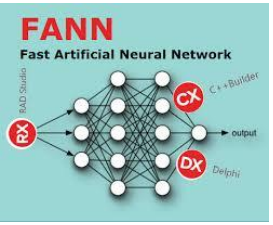
Цель: обеспечить кластеризацию распределённых данных обучив нейронную сеть

Задачи:

1. Провести обзор подходов к решению задачи
2. Рассмотреть методы распределения вычислений
3. Создать и обучить нейронную сеть локально
4. Применив SGD обучить сеть распределённо
5. Определить предел возможного ускорения вычислений при увеличении количества вычислителей

Использованные технологии и инструментарий

Python – богатый ЯП применяемый для машинного обучения



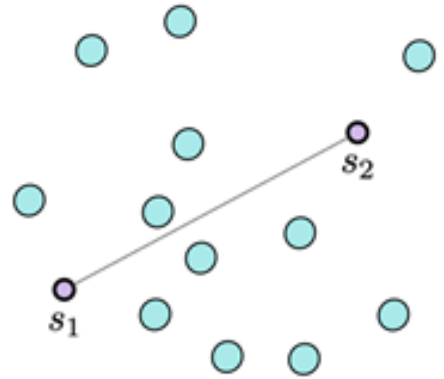
Fast Artificial Neural Network – библиотека машинного обучения, для работы достаточно указать гиперпараметры и набор данных

SimpleHTTPServer – модуль для обеспечения лёгкого создания HTTP серверов

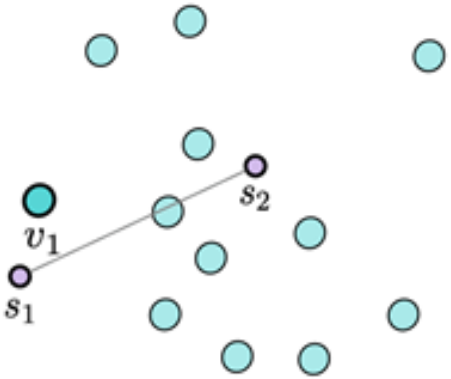
Особенности реализации

- Данная работа представляет из себя клиент-серверное решение
- Работа пяти HTTP серверов обеспечивается через TCP/IP протокол, каждый из которых запущен локально и отводится под отдельные задачи
- Наборы данных представленные в работе и используемые для обучения и тестирования взяты с https://www.kaggle.com/datasets
- Использование ASGD, и модели параллелизма для обеспечения распределённого обучения

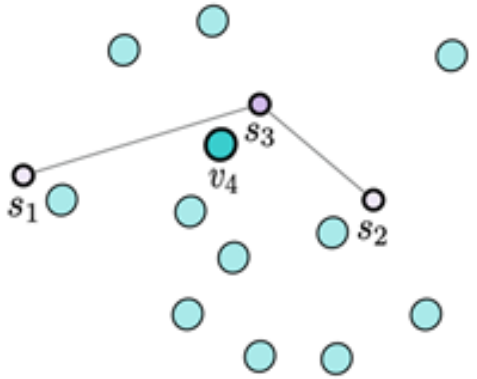
Принцип работы Растущего нейронного газа



1) Создать два случайных нейрона и соединить их. Это начало графа GNG.



2) Выбрать точку данных v1 и двигать ближайший нейрон(в данном случае s1), так же как тот с которым он непосредственно соединён(здесь s2) ближе к v1



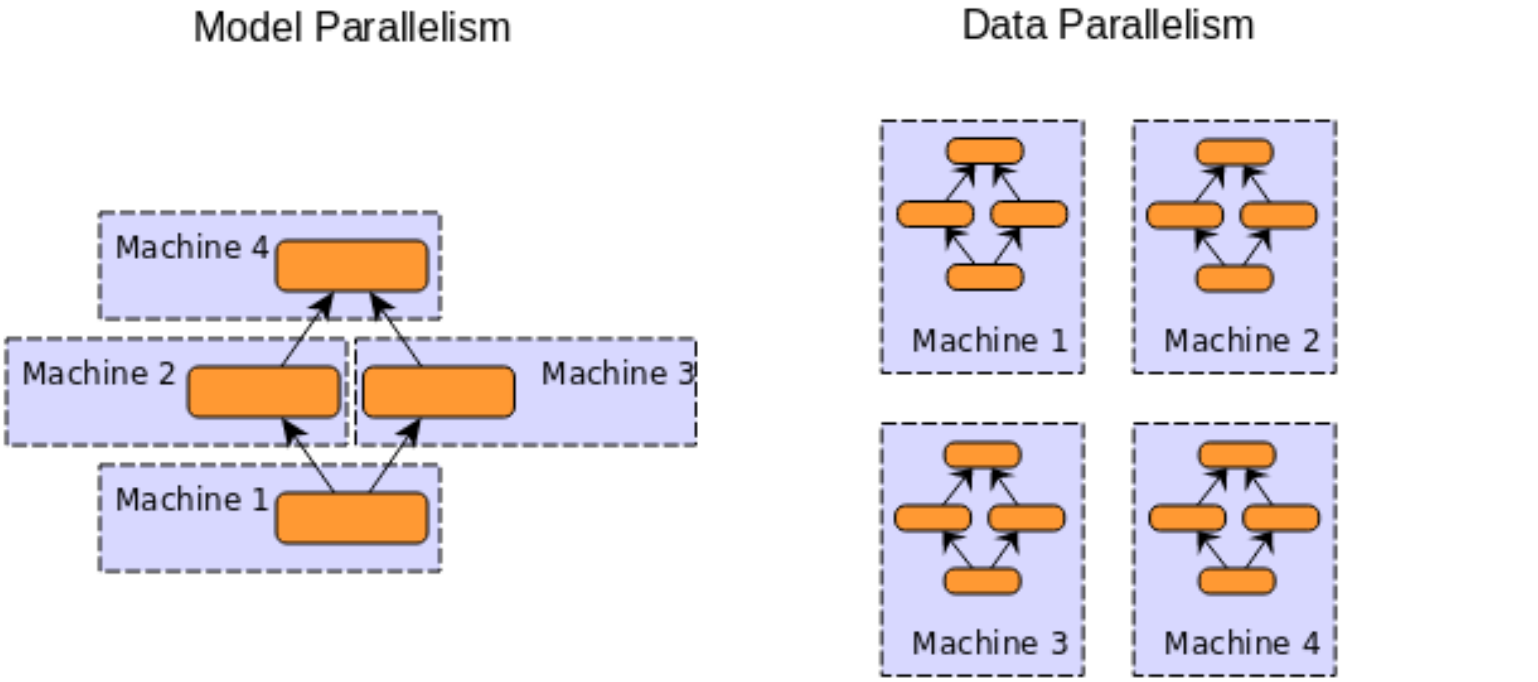
3) После данных шагов создать нейрон между нейроном с наибольшей ошибкой и нейроном соединённым с нейроном с наибольшей ошибкой

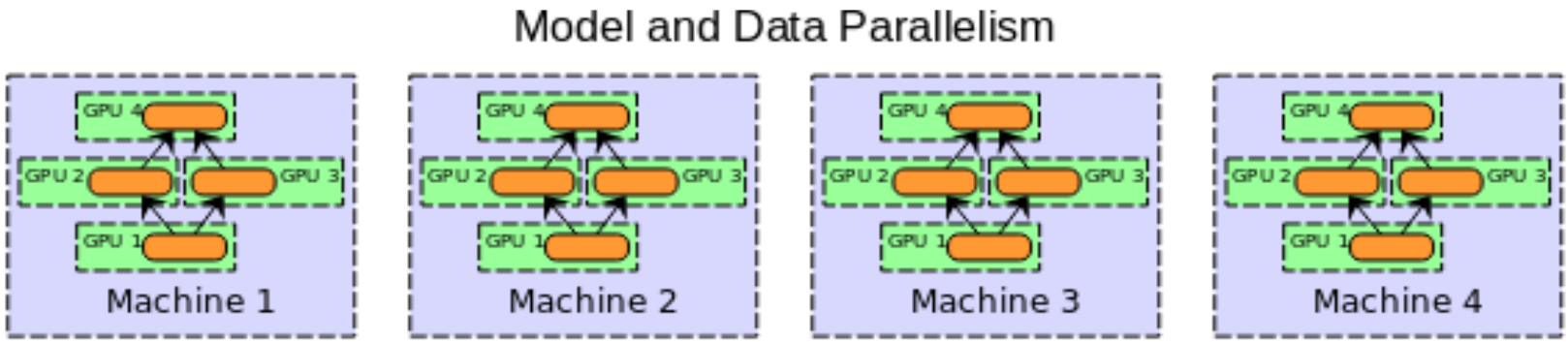
Основная идея работы GNG

Рассмотрение аналогов

Имя	**Основа сети**	**Ограничение на обработку**	**Работа с шумом и т.д**	**Необходимость спецификации**
partSOM	K-means/ SOM	До десяти тысяч экземпляров	Нет	Да
DPDC	DBSCAN	До ста тысяч экземпляров	Да	Нет
NNDDC	GNG	Более ста тысяч экземпляров	Да	Нет

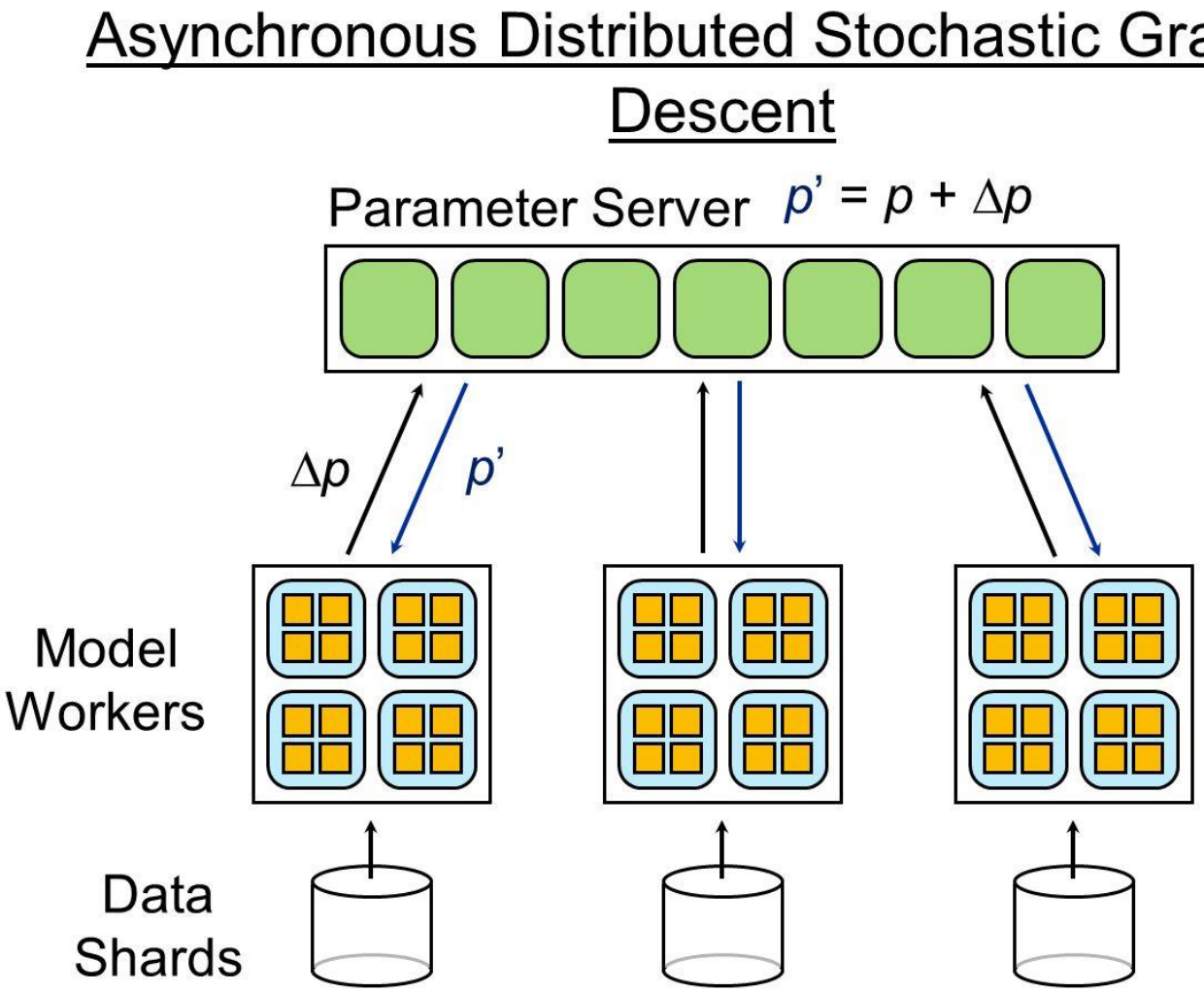
Методы распределения вычислений





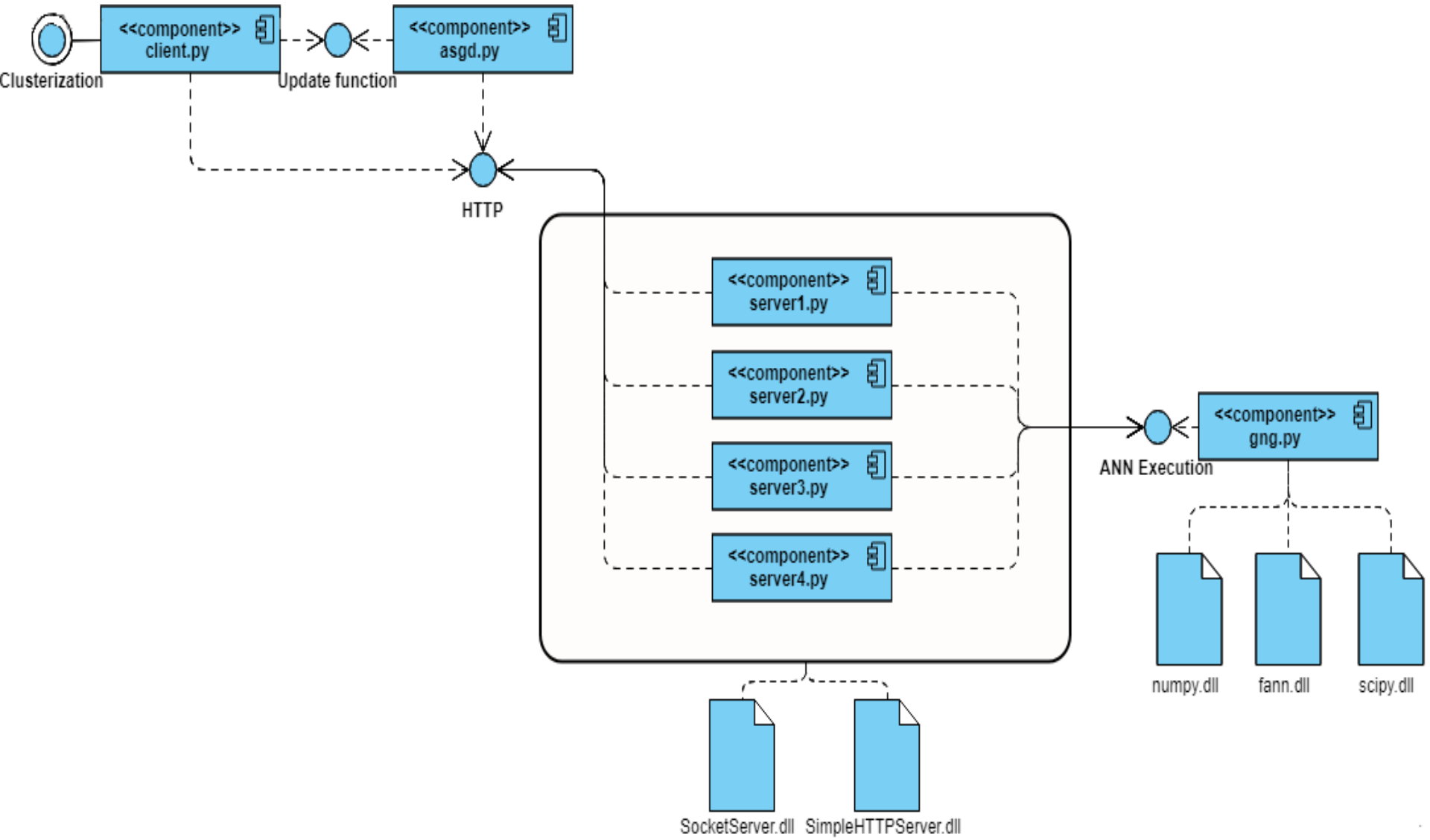
Представление методов параллелизм данных и параллельности модели и их объединения

ASGD для обновления параметров сети



Был реализован асинхронный стохастический градиентный спуск, как оптимизирующая функция для усреднения параметров сети

Архитектура проекта



Оценки времени вычислений и стоимости связи

1 Computation time

$$O(\frac{p}{q}logk) + O(plog\frac{1}{e}) \quad (1)$$

2 Communication cost

$$O(Nk) + O(pk) \quad (2)$$

3 Communication time

$$O(plogk) \quad (3)$$

Где p – размер параметров, e – желаемая ошибка, k – количество машин, q – количество процессоров, N – количество экземпляров в наборе данных

Результаты при тестировании

Наименование набора данных	**Размер наборов/ параметров**	**Время необходимое на обучение**	**Точность при тестировании**
Diabetic Disease	384/8	2-3 часов	94,3
Market Basket	90/4	20-30 минут	85,1
Wine Reviews	130000/7	1-2 дней	92,2

Ограничение распределённых систем

Согласно закону Амдала и вытекающего из него закона Густавсона - Барсиса, мы встречаемся с невозможностью преодоление определённого ускорения системы, предел для проводимого исследования в рамках ВКР оказался трёхкратным для 16 процессоров

Закон Густафсона — Барсиса выражается формулой:

$S = s + (1-s)n = n + (1-n)s$, где

- s — доля последовательных расчётов в программе,
- n — количество процессоров.

Апробация работы

- Репозиторий проекта https://github.org/Bedrang/NNDDC.

Заключение

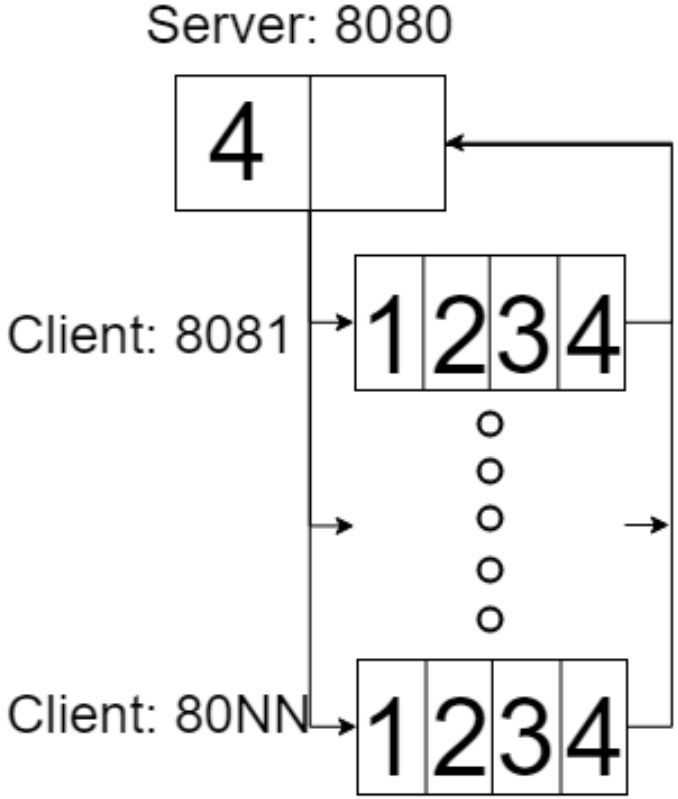
- Проделанный обзор методов показал необходимость разработки данного алгоритма с применением SGD
- Сформулированы критерии которые необходимо определить для получения точных результатов
- Экспериментальное исследование скорости работы алгоритма показало, что прирост производительности обучения сети возрос почти в два раза
- Дальнейшие направления исследований включают в себя доработку метода усреднения параметров, и использования модификаций GNG(IGNG)

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

Спасибо за внимание!

Выполнил: Итальянцев Ян Викторович, гр. 3304
Руководитель: Борисенко Константин Алексеевич, к.т.н.
ассистент кафедры МОЭВМ

Общая схема работы системы



Принимает вектора вычислений и оптимизирует параметры

Такое количество дополнительных вычислительных узлов сколько необходимо

1 - Входной слой сети

2 - Скрытый слой сети

3 - Выходной слой сети

4 - стохастический градиентный спуск, оптимизирующая функция