Задачу нахождения формулы, которая описывает некую зависимость, называют символьной регрессией. Более формально, её можно сформулировать как построение регрессионной модели в виде суперпозиции заданных функций.

Аналоги. Символьная регрессия отличается от традиционной линейной, квадратичной или полиномиальной регрессии, которые просто находят числовые коэффициенты для функции, вид которой заранее известен.

Актуальность выбранной темы обосновывается тем, что символьная регрессия может использоваться для получения эмпирических зависимостей на основе экспериментальных данных. Она широко применяется для решения задач моделирования и прогнозирования.

Проблема. Существует два набора данных. Есть предположение, что они зависят друг от друга. Необходимо найти вид функциональной зависимости.

Объектом исследования выступает алгоритм установления функциональной зависимости.

Целью выпускной квалификационной работы является нахождение и отработка методики восстановления функциональной зависимости данных.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1. Изучить соответствующую литературу.
2. Разработать алгоритм нахождения функциональной зависимости.
3. Реализовать программу по данному алгоритму.
4. Протестировать программу с реальными данными.
5. Обобщить полученные результаты и сделать соответствующие выводы.

Решать задачу символьной регрессии мы будем с помощью генетического программирования. Генетический алгоритм использует механизмы естественной эволюции, основанные на следующих принципах: естественный отбор, скрещивание и мутация.

Генетический алгоритм состоит из следующих этапов:

* Создание исходной популяции.
* Оценка пригодности особей.
* Выбор родителей для скрещивания.
* Создание потомков выбранных пар родителей.
* Мутация особей.
* Проверка достижения результата.

Особи популяции представлены в виде деревьев, которые в свою очередь представляют определённую функцию. Внутренние узлы дерева выбираются из функционального множества, а листья дерева из терминального, состоящего из констант и переменных.

В качестве меры пригодности используется квадратичная ошибка. Мы вычисляем ее значение для каждой особи и впоследствии нормализуем. Выбор родителей происходит пропорционально нормализованным значениям пригодности.

Скрещивание происходит следующим образом.

1. Случайно выбираются родители и внутренний узел у каждого из них. Далее необходимо убедиться, что выбранные узлы взаимозаменяемы, т.е. у них совпадает арность операций. Иначе, во втором дереве выбирается другой узел с последующей проверкой.
2. Затем производится обмен поддеревьев, которые определены этими узлами.
3. Вычисляется размер ожидаемых потомков. Если он не превышает заданный порог, такой обмен ветвями запоминается.

После выполнения кроссинговера с заданной малой вероятностью происходит мутация. Мутация бывает трех видов: узловая, усекающая и растущая.

* Узловая мутация состоит в замене терминального символа на другой терм или в замене функции на другую функцию.
* Усекающая мутация представлена в виде замены функции на терминальный символ, а растущая в виде замены терминального символа на функцию.

Результаты. Была разработана программа, устанавливающая функциональную зависимость данных посредством генетического программирования. Для проверки ее работы был проведен эксперимент. Задана исходная функция, выборка данных и максимальная квадратичная ошибка, равная 1. В итоге была выбрана следующая функция.

Результаты работы программы можно продемонстрировать в виде графика, на котором синим цветом изображена исходная функция, оранжевым – полученная. Точки представляют значения множества выборки.

Видно, что полученная функция достаточно точно совпадает с заданной функцией в точках выборки данных, т.е. ведет себя как тестовая функция.