ГЕНЕТИЧЕСКОЕ ПРОГРАММИРОВАНИЕ КАК МЕТОД СЖАТИЯ ДАННЫХ

Бойко А.В., группа АСУ00а
Руководитель доц. Орлов Ю.К.

Вопрос экономного кодирования информации в системах управления был поставлен в первой половине 1970-х годов, но не потерял актуальности и до сих пор. В связи с прогрессом компьютерных и инженерных разработок поток информации, получаемый из различных областей науки и техники возрос в сотни и тысячи раз.

Для хранения и автоматизации обобщения информации создаются специальные банки данных. В странах с небольшой по площади территорией принято создавать единые банки для всех компонентов гидросферы. Такой огромный объем информации, который нужно ежедневно обрабатывать (а главная проблема состоит в том, что их нужно хранить в течение несколько десятков лет для дальнейшего анализа) приводит к необходимости сжатия информации и дальнейшем ее восстановлении с минимальными потерями.

На сегодняшний день существует множество алгоритмов сжатия, которые имеют как свои достоинства, так и недостатки. Основными из них являются сжатие таблиц (экономия места достигается за счет удаления избыточных копий значений данных в таблице), упаковка битов (заключается в кодировании значения атрибута битовыми последовательностями фиксированной длины), арифметическое кодирование (предполагаемая требуемая последовательность символов рассматривается как некоторая двоичная дробь из интервала [0, 1). Результат сжатия представляется как последовательность двоичных цифр из записи этой дроби) и др. Главными недостатками всех методов является то, что при сжатии происходит значительная потеря информации, которую нельзя

восстановить. Обычно приводимые оценки степени сжатия баз данных составляют 2..5 раз.

В качестве метода сжатия гидрологических данных в данной статье предлагается метод генетического программирования. Метод генетического программирования строится на основании концепции генетических алгоритмов. По этой концепции случайным методом генерируются 2 особи, которые на следующем шаге используются для репродукции новых пар, а затем генерации новых особей из этих пар. Новые особи подвергаются отбору и мутации, а затем, исходя из постановки задачи, выбираются новые особи, которые могут породить более оптимальное потомство. В генетическом программировании хромосомами являются программы. Программы представлены как деревья с функциональными (промежуточные) и терминальными (конечные) элементами. Терминальными элементами являются константы, действия и функции без аргументов, функциональными - функции, использующие аргументы.

Набор терминалов, представляет собой компоненты, из которых будет создаваться функция для полного или частичного решения проблемы. Второй шаг заключается в определении функционального множества, элементы которого должны использоваться для генерации математических выражений. Рассмотрим для примера функцию $\sqrt{arctg(x+3)*\ln 5}$. Ее можно представить в следующем виде:

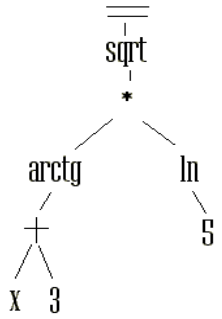

Рисунок 1 – Функция $\sqrt{arctg(x+3)*\ln 5}$, представленная в виде дерева

Для данного дерева терминалами будут являться – x, 3, 5, а функционалами – arctg, sqrt и ln. Рассмотрим 2 функцию, а затем взимодействие их:

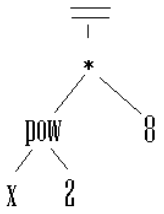

Рисунок 2 – Функция $8x^2$, представленная в виде дерева

Если теперь подвергнуть 2 этих особи (дерева) операции кроссинговера, то в результате получим 2 новых особи:

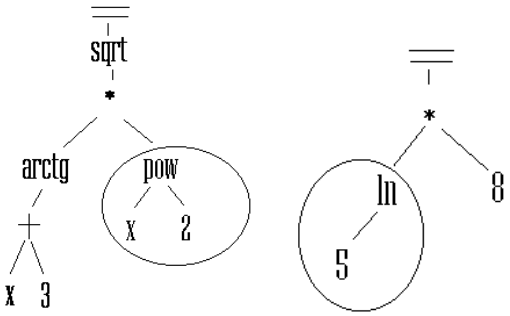

Рисунок 3 – Генерация новых особей

Из приведенного рисунка видно, что деревья обменялись составными частями, в результате чего возникли новые функции $\sqrt{arctg(x+3)*x}^2$ и 5*ln8, которая является константой.

Допустим, в течение 5 лет производится измерение скорости течения «Гольфстрим» на разной глубине. За один день производится 1000 измерений. Тогда метод сжатия будет заключаться в подборе методом генетического программирования (исходя из начальных функционалов и терминалов) 5-6 или более функций, с помощью объединения которых можно будет потом получить

весь набор точек. Возможно, что скорость течения будет одинакова на всех точках, тогда функция будет одна и представлять собой константу. Также возможна ситуация, когда на какой-то точке измерения будет наблюдаться скачок. При этом не получится подобрать никакую функцию и придется хранить эту точку отдельно. В любом случае данных нужно будет хранить намного меньше.

Кроме самого результата сжатия такой метод предполагает одновременный анализ данных, который может прослеживаться на больших временных отрезках и при этом не обязательно раскодирование. Так например при анализе температуры, видя, что функция температуры имеет приблизительный вид прямой, можно сократить измерения на данном участке объекта или сделать их более дискретными.

Разбивая задачу сжатия данных таким методом можно выделить следующие этапы:

1. Задаться набором минифункций, на основании которых будет производиться поиск функции или набора функций, кодирующих заданную последовательность параметров.

Такая последовательность может быть как определена пользователем в диалоге, так и использована непосредственно в алгоритме (стандартные наборы типа - +,-,*,/, $\sqrt{}$,cos, sin, log и т.д)

2. Определить диапазон, в котором будут находиться возможные терминалы.

Вариантом поиска коэффициентов является метод наименьших квадратов, с помощью которого можно будет решить систему уравнений для поиска коэффициентов. Кодирование функцией будет фактически означать ее аппроксимирование. В точках набора сгенерированная функция будет отличаться от экспериментального значения на некоторую величину $\xi=f(t_i)-h_i$. Чем меньше эта величина, тем более точной является аппроксимирующая функция. Таким образом для популяции значение величины ξ будет являться

значением фитнесс-функции. Применяя метод наименьших квадратов мы одновременно получим коэффициенты искомой функции.

3. Определить параметры алгоритма, такие как способ мутации, способ селекции.

Способ мутации заключается в произвольном выборе узла. Селекция обычно проводится методом рулетки, т.е. определяются вероятности выживаемости особей и выбираются те, у которых эта веротяность больше. После селекции “плохие” особи отбрасываются для того, чтобы не произошло “затухания потомства”.

4. Провести анализ алгоритма на разных наборах данных при различной точности.

5. Определить недостатки и достоинства данного метода.

Два последних этапа носят сугубо аналитический характер. Главной их задачей будет доказать, что используемый метод работает на больших выборках. Минимальный коэффициент сжатия данных планируется 5-7 единиц, что должно превысить существующие на данный момент коэффициенты среди методов сжатия числовой информации.

Перечень ссылок

1. Методы сжатия данных/ http://compression.ru
2. А.И. Змитрович Интеллектуальные информационные системы. - Минск.: НТООО "Тетра Системс", 1997. - 368с.