УДК 004.85

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ ПОТОКА РАБОТ ДЛЯ КЛАССИФИКАЦИИ С ПОМОЩЬЮ ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Пендряк А.А., магистрант гр. М4138, Университет ИТМО, Санкт-Петербург.

Научный руководитель - Фильченков А.А., к.ф.-м.н., доцент каф. КТ, Университет ИТМО, Санкт-Петербург.

На сегодняшний день классификация является наиболее частой и востребованной задачей в машинном обучении. Существует огромное число задач, для решения которых применяется классификация, а также существует огромное количество алгоритмов классификации. Такое количество различных алгоритмов обусловлено не только разнообразием решаемых задача, но еще и тем, что не существует оптимального алгоритма для всех типов данных (No Free Lunch Theorem). Поэтому при решении очередной задачи мы вынуждены искать алгоритм, наиболее подходящий для заданного набора данных. Подобный поиск значительно осложняется еще и тем, что при решении задач подобного рода требуется выбрать не только подходящий алгоритм классификации, но также и алгоритм для предобработки данных и подобрать их параметры. В результате ручной поиск подходящего сочетания алгоритмов и их параметров очень трудоемок. В связи с этим возникает потребность в автоматическом подборе оптимальных алгоритмов и параметров для решения задач классификации. Стоит сразу отметить, что такой подбор должен значительно отличатся от перебора вариантов, так как количество различных комбинаций алгоритмов и их параметров очень велико.

Целью данной работы является исследование способов автоматического построения потока работ для классификации и улучшение эффективности их работы.

Поток работ для задачи классификации – это направленный ациклический граф, в узлах которого содержатся базовые алгоритмы, такие как классификаторы, алгоритмы предварительной обработки (feature selection и feature extraction), разветвители, преобразующие входящие данные в несколько выходящих потоков, и агрегаторы, которые объединяют результаты нескольких классификаторов.

Как показывает практика, для решения задач классификации не требуется придумывать одну очень сложную модель данных. Один из самых эффективных способов решения данной задачи – агрегирование результатов нескольких простых моделей. Поэтому в потоке работ могут присутствовать более одного классификатора и агрегаторы, которые будут объединять их результаты.

Так как функция эффективности (точность, f-мера или любая другая метрика эффективности классификации, наиболее подходящая в решаемой задаче) нетривиальным образом зависит от структуры потока работ, то для оптимизации данной функции можно применить метод генетического программирования. В качестве особей в популяции будет выступать сам поток работ с вектором значений параметров алгоритмов, из которых он построен, а функция эффективность будет в данном случае функцией приспособления. Операцию мутации можно осуществить как замену случайного подграфа на другой случайный подграф, подходящий по типу данных, а операцию скрещивания – как обмен случайными подграфами двух потоков работ.

Описанный выше подход предложен в статье «Evolving Workflow Graphs Using Typed Genetic Programming». Недостатками данной работы является маленькое количество наборов данных, на которых демонстрируются результаты, и сильно ограниченное множество базовых алгоритмов.

	auto	openml.org
iris	0.98	0.98
hayes-roth (binary)	0.8549	0.8636
rabe_266 (binary)	0.9769	0.9833
tae	0.7438	0.8079

Сравнение точности классификации на некоторых наборах данных с лучшим результатом с ресурса по машинному обучению openml.org.

Первые эксперименты показывают, что данный подход строит довольно эффективные потоки работ для некоторых наборов данных. Но для других данных модель получается не очень эффективная. Возможно, при их построении стоит строить потоки работ с большим количеством узлов и увеличить максимальное количество особей в популяции и количество самих популяций.

Другим недостатком данного подхода является продолжительная работа алгоритма генетического программирования, особенно на больших наборах данных.

В настоящий момент разрабатываются решения этих недостатков в основных двух направлениях:

- Расширение списка базовых алгоритмов, в особенности алгоритмов предобработки и агрегации;
- Инициализация первой популяции при помощи мета-обучения. В предположении, что для близких наборов данных алгоритмы работают похожим образом, можно инициализировать первую популяцию не только случайными особями, но и лучшими результатами для близких наборов данных, тем самым значительно уменьшить количество популяций для получения оптимального результата.

Магистрант гр. М4138 _____ Пендряк А.А.

Научный руководитель _____ Фильченков А.А.

Зав. кафедрой КТ, Университет ИТМО _____ Васильев В.Н.