ГЕНЕРАЦИЯ ЭКЗЕМПЛЯРОВ ДЛЯ ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ ПО ИХ ХАРАКТЕРИСТИЧЕСКОМУ ОПИСАНИЮ[1]

Забашта А. С., магистрант кафедры компьютерных технологий Университета ИТМО, azabashta@corp.ifmo.ru
Фильченков А. А., к.ф.–м.н., доцент кафедры компьютерных технологий Университета ИТМО, aaafil@mail.ru

Аннотация

В данной работе предложен метод генерации экземпляров для задачи классификации по заданному характеристическому описанию. Данный метод основан на генетическом алгоритме, для которого были разработаны операторы мутации и кроссовера экземпляров на основе удаления или добавления объектов и признаков в набор данных. Предложенный метод был протестирован в нетривиальном трёхмерно мета–признаковом пространстве и оказался в 2,7 раза точнее наивного метода, который не учитывает требуемое характеристическое описание.

Введение

В данной работе рассматривается одна из самых популярных и востребованных задач машинного обучения — задача классификации. На сегодняшний день существует большой набор различных алгоритмов классификации, но среди них нет одного универсального алгоритма [1]. Если на одной модели данных конкретный алгоритм работает точнее остальных классификаторов, то всё равно найдётся такая модель данных, на которой он будет уступать другим алгоритмам. Таким образом, при исследовании классификаторов следует рассматривать не все возможные экземпляры, а некоторые области пространства.

Для работы с пространством задач классификации и определением областей в нём исследователи используют общие признаки, которыми обладают все возможные экземпляры для задачи классификации — их численные характеристики, которые и определяют конкретное пространство [2]. При работе с полученными пространствами используется предположение, что на достаточно близких в пространстве

[1] Работа выполнена при финансовой поддержке правительства Российской Федерации, грант 074-U01 и РФФИ, грант 16-37-60115-мол_а_дк.

наборах данных, алгоритмы работают похожим образом.

Проблема в том, что пространство может быть недостаточно заполнено для корректной работы с ним. Например, в пространстве могут быть достаточно большие пустоты, в которых не нашлось экземпляры из прообраза отображения характеристик. Также могут существовать экземпляры, которые будут близкими к существующим, но результаты работы алгоритмов на них будут сильно различаться.

Цель данной работы — разработка методов заполнения характеристического пространства, которые генерируют по характеристическому описанию экземпляры для классификации, которые максимально приближены к исходному описанию.

Метод

Задача классификации

Задача классификации относится к классу задач обучения с учителем [3]. Традиционно экземпляр задачи классификации представляется в виде матрицы «Объекты–признаки». Каждая её строка — определённый объект, который описывается вектором признаков, характеризующих его свойства, и меткой класса, к которому он принадлежит. В данной работе рассматриваются только экземпляры с вещественными признаками и бинарными классами $\{p, n\}$.

Для исследования классификаторов в рамках мета–обучения требуется характеристическое описание наборов данных – вектор мета–признаков, которые описывают свойства всех возможных экземпляров задачи классификации. Из экземпляра можно выделить тривиальные характеристики: число признаков, число объектов и число классов. Достаточно просто сгенерировать набор данных с соответствующими характеристиками. Гораздо труднее обращать другие характеристики, например статистические и теоретико–информационные метрики признаков и классов или состояние дерева принятия решений, построенном на данном экземпляре (статистика числа листьев, ветвей, глубины или ширины).

Описание предлагаемого подхода

Генерацию экземпляров для классификации по их характеристическому описанию можно представить, как минимизацию расстояния от характеристического описания текущего экземпляра до требуемого. Для решения данной задаче использовался один из алгоритмов эволюционного вычисления — генетический алгоритм [4], так как он

способен работать с абстрактными представителями популяций. Для его работы требуется начальная популяция, оператор кроссовера и оператор мутации.

Оператор мутации был задан как функция, добавляющая или удаляющая случайное множество признаков и объектов из мутирующего набора данных. Так как экземпляр для задачи классификации можно представить как матрицу, удаление или добавление признаков и объектов можно представить как удаление или добавление соответствующих строк и столбцов. После добавления новой строки – объекта, его признаки заполнялись значениями соответствующих признаков уже существующих случайно выбранных объектов того же класса. После добавления столбца создавалась случайная функция от множества существующих признаков в вещественные числа и применялась ко всем объектам.

Оператор кроссовера заключался в объединение случайных пар объектов равных классов с предварительным уравниванием размеров, описанным ранее оператором добавления и удаления объектов. После этого полученный большой экземпляр разбивался по случайному подмножеству признаков на два меньших, которые являются результатом кроссовера. Пример кроссовра представлен на рисунке 1.

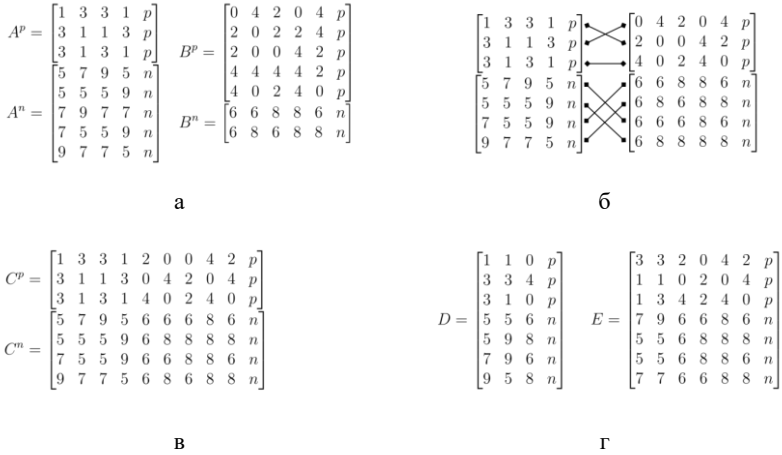

Рисунок 1: Пример кроссовера экземпляров *A* и *B* в *D* и *E*.

Стоит отметить, что описанные выше операторы являются базовыми для алгоритмов оптимизации основанных на эволюционных вычислениях.

Например, мутацию можно применить в алгоритме имитации отжига [5].

Эксперимент

Описание

Для сравнительной проверки предложенного метода на основе генетического алгоритма был использован наивный метод генерации экземпляров не учитывающий требуемое характеристическое описание. Этот метод генерировал экземпляры из почти пустой матрицы, используя оператор добавления признаков. Число объектов и признаков для нового экземпляра выбиралось равновероятно из отрезка $(0; 200)$. Работа обоих методов была ограничена равным числом просмотренных экземпляров: 6500 штуками. Проверка проводилась в трёхмерном характеристическом пространстве, образованном из нетривиальных мета–признаков: средняя попарная корреляция признаков (C), средняя взаимная информация признаков с классом (I) и среднее число ветвей в дереве принятия решений, построенном на данном экземпляре (B).

В качестве погрешности использовалось евклидово расстояние от характеристик полученного экземпляра до заданного описания. Для корректного подсчёта расстояний значения по каждой координате нормировались на среднеквадратичное отклонение и сдвигались на математическое ожидание. Для нормирования характеристического пространства, а также в качестве начальной популяции для генетического алгоритма использовались 600 существующих экземпляров для задачи бинарной классификации с ресурса OpenML.org. В качестве искомой точки была выбрана точка начала нормализованных координат.

Результаты

Результаты работы описанных ранее алгоритмов представлены на рисунках 2–4. Как видно по результатам работы, алгоритму удалось найти достаточно близкую к заданному описанию точку и достаточно компактно заполнить пространство вокруг неё. Подход с использованием генетического алгоритма оказался в 2,7 раза точнее, чем простая генерация данных, которая не учитывают требуемое характеристическое описание.

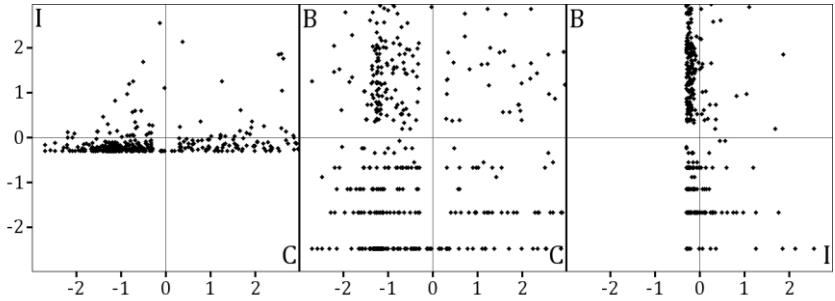

Рисунок 2: Проекции трёхмерного нормализованного пространства с распределением существующих экземпляров для задачи бинарной классификации. Ближайшая точка до центра находится на расстоянии 0,550.

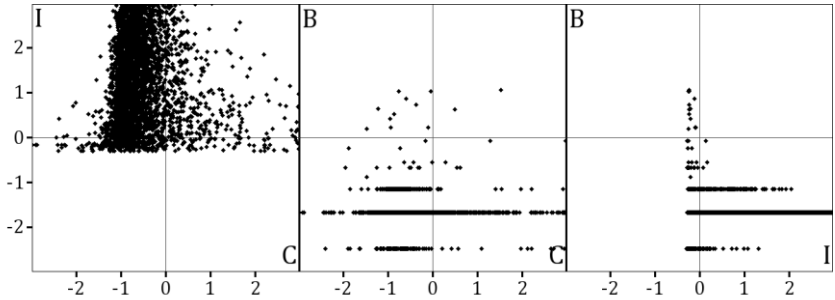

Рисунок 3: Проекции трёхмерного нормализованного пространства с распределением экземпляров полученных наивной генерацией. Ближайшая точка до центра находится на расстоянии 0,228.

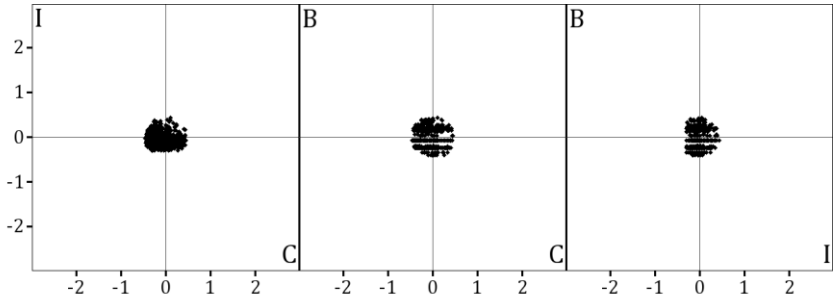

Рисунок 4: Проекции трёхмерного нормализованного пространства с распределением экземпляров полученных на последней эпохе генетического алгоритма. Ближайшая точка до центра находится на расстоянии 0,084.

Заключение

В данной работе предложен метод генерации экземпляров данных для задачи классификации по характеристическому описанию. Данный метод основан на генетическом алгоритме, для которого были разработаны операторы мутации и кроссовера экземпляров на основе удаления или добавления объектов и признаков в набор данных. Предложенный метод был протестирован в нетривиальном трёхмерно мета–признаковом пространстве и оказался в 2,7 раза точнее наивного метода, который не учитывает требуемое характеристическое описание.

Реализованный подход можно применять в любых сферах, связанных с характеристическим описанием данных, для повышения точности алгоритмов, работающих с ними. К таким сферам относятся задачи сравнения алгоритмов, а также предсказание наилучшего алгоритма. Предложенный метод можно использовать с любыми вещественными характеристиками данных, но он легко может быть модифицирован под другие пространства с заданной метрикой. В алгоритме могут использоваться пространства любой размерности, а реализованные операторы применяться в других алгоритмах эволюционного вычисления.

Литература

1. Wolpert D.H., Macready W.G. No free lunch theorems for optimization // IEEE transactions on evolutionary computation. 1997. Vol. 1(1). P. 67–82.
2. Vilalta R., Drissi Y. A perspective view and survey of meta–learning // Artificial Intelligence Review. 2002. Vol. 18(2). P. 77–95.
3. Николенко С.И., Тулупьев А.Л. Самообучающиеся системы. М.: МЦНМО, 2009. 287 с.
4. Скобцов Ю.А. Основы эволюционных вычислений. Донецк: ДонНТУ, 2008. 326 с.
5. Kirkpatrick S. et al. Optimization by simulated annealing // Science. 1983. Vol. 220(4598). P. 671–680.