УДК 004.85

Исследование отсекающих правил ансамблей фильтрующих алгоритмов выбора признаков

Красноцветов В. В., Университет ИТМО, Санкт-Петербург

Научный руководитель – Сметанников И. Б., Университет ИТМО, Санкт-Петербург, к.т.н., научный сотрудник ФИТиП

Введение

Машинное обучение – подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных выявлять скрытые в данных зависимости. Нередко, в задачах биоинформатики встречаются большие объемы данных, но не все данные одинаковы полезны. Каждый экземпляр данных может характеризоваться набором признаков. Благодаря машинному обучению наиболее релевантные признаки могут быть выделены среди всего множества признаков.

Существует множество различных методик для получения данного результата ([1], [2]), в частности одни из наиболее известных: фильтрующие методы, методы-обертки и встроенные методы. Среди фильтрующих методов выделяются ранжирующие фильтры: зависимости между признаками не учитываются, сами признаки сортируется по некоторой оценки и выбирается некоторое количество наиболее ценных признаков.

MeLiF [3] – современный ансамблевый алгоритм отбора признаков. Результатом работы алгоритма является оптимальная линейная комбинация базовых ранжирующих фильтров, позволяющая отобрать такое подмножество признаков исходного набора данных, которое максимизирует метрику классификации.

Цель

Целью данной работы является улучшение качества работы алгоритма *MeLiF*, а так же оптимизировать время работы данного алгоритма в задачах биоинформатики.

Базовые положения исследования

В процессе работы алгоритм *MeLiF* совершает обход линейного пространства метрик с помощью покоординатного спуска по фиксированной сетке и жадно выбирает координату и направление движения по ней. В силу того, что для обхода использован метод покоординатного спуска по фиксированной сетке, алгоритм очень чувствителен к тому, какая будет выбрана начальная точка и каким именно будет шаг сетки.

Другая проблема алгоритма *MeLiF*, скорость работы. Алгоритм *MeLiF* обходит гиперпространство размера N, где N – количество функций оценки признака. В процессе обхода пространства алгоритм *MeLiF* оптимизирует линейную комбинацию весов функций оценок, а после чего, как и любой ранжирующий фильтр, применяет правило отсечения. Благодаря данной реализации алгоритм *MeLiF* может несколько раз посещать точки пространства, из которых будет выбран один и тот же набор наиболее релевантных признаков.

Рассмотрим рисунок 1, на котором приведен процесс изменения релевантных признаков при двух функциях оценок каждого признака. Можно заметить, что при обходе данного пространства, набор наиболее релевантных признаков изменяется после пересечения линий признаков. Красная пунктирная линия показывает области, на которых конечный набор наиболее релевантных признаков не изменяется.

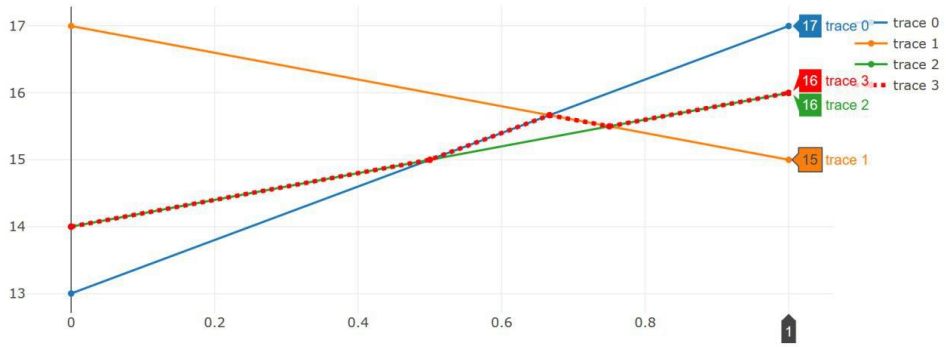

Рисунок 1 - Изменения набора релевантных признаков

Основная задача разработанного алгоритма – отыскать такие области. Для нахождения таких областей в N-ом пространстве будет использоваться следующий алгоритм:

1. Каждая гиперплоскость, соответствующая функции оценки признака, вписывается в исходное гиперпространство в результате чего получается выпуклая фигура.
2. Полученная выпуклая фигура разрезается множеством гиперплоскостей, соответствующих функциям оценки других признаков. В результате разрезания фигуры другой гиперплоскостью, получается две выпуклых фигуры. При этом той части исходной фигуры, что оказывается под гиперплоскостью увеличивается счетчик.
3. Из множества полученных выпуклых фигур выбираются только те, счетчик которых удовлетворяет отсекающему правилу.
4. Для каждой выбранной фигуры вычисляется внутренняя точка и проецируется на пространство обхода. Данная точка добавляется в ответ.

Результаты

В ходе данной работы был разработан алгоритм, который улучшает результаты и скорость работы алгоритма *MeLiF*, а также был реализован прототип данного алгоритма.

Список литературы

1. Guyon I., Elisseeff A. An introduction to variable and feature selection //Journal of machine learning research. – 2003. – Т. 3. – №. Mar. – С. 1157-1182.
2. Dash M., Liu H. Feature selection for classification //Intelligent data analysis. – 1997. – Т. 1. – №. 1-4. – С. 131-156.
3. Smetannikov I., Filchenkov A. MeLiF: filter ensemble learning algorithm for gene selection //Advanced Science Letters. – 2016. – Т. 22. – №. 10. – С. 2982-2986.