УДК 004.02

ОБЗОР ТЕКУЩЕГО СОСТОЯНИЯ МУЛЬТИВАРИАТИВНЫХ МЕТОДОВ И АЛГОРИТМОВ ОТБОРА ПРИЗНАКОВ, ОСНОВАННЫХ НА АНСАМЛЯХ

Глухов В.Н. (Университет ИТМО)
Научный руководитель – доцент, к.т.н. Сметанников И.Б.
(Университет ИТМО)

Аннотация. Исследование представляет собой систематизацию знаний и положений о существующих подходах к проблеме отбора признаков, основанных на объединении их мультивариативных мер значимости в виде ансамблевых моделей, хорошо зарекомендовавших себя в задаче обучения моделей машинного обучения. Ансамблевые модели являются более точными чем отдельные модели ввиду усреднения ответов нескольких моделей, что позволяет устранить слабые стороны каждой из них.

Введение. Проблема отбора признаков в задаче предобработки данных для обучения моделей машинного обучения является одной из основных ввиду её важности. Банки, ритейл, интернет-компании накопили терабайты данных о пользователях, чтобы проводить различного рода аналитику, но такая аналитика становится затруднительной ввиду больших измерений данных, имеющихся, например, для описания каждого пользователя, что прямым образом сказывается на производительности и временных затратах на обучение моделей. Хотя задача производительности решается увеличением мощности аппаратных ресурсов, принцип “garbage in – garbage out” все еще имеет место быть, потому что не все из имеющихся данных имеют какое-то объясняющее и значимое значение для решаемой задачи.

Хотя и существует большое число различных методов и алгоритмов решения проблемы отбора признаков, все они в большинстве своем разнородны, имеют разные предпосылки, определены в различных предметных областях и подобластях таких дисциплин как информационная теория и статистическое обучение. Особенно популярным направлением разработки методов снижения размерности данных является ансамблирование, хорошо зарекомендовавшее себя в задачах машинного обучения и в большинстве случаев показывающее лучшее качество нежели отдельные алгоритмы.

Исследование ставит своей целью систематизацию информации о текущем состоянии области мультивариативных алгоритмов отбора признаков, основанных на ансамблях.

Основная часть. В ходе исследования проанализированы возможные постановки задачи отбора признаков на основе ансамблей и выделены три наиболее широкие группы.

1) Ансамбли, основанные на моделях, представляют из себя обучение различных классификаторов на разных подмножествах исходных данных. Таким образом, на выходе ансамбля имеется несколько метрик качества моделей, по которым по некоторму усредняющему правилу определяют отобранные признаки.

2) Ансамбли, основанные на рангах, представляют из себя запуск нескольких простых ранговых алгоритмов на исходных данных по различному ранговому правилу. На выходе ансамбля имеется множество признаков с рангами, основанные на различных ранговых правилах, которые усредняются, и отбираются наилучшие признаки.

3) Ансамбли, основанные на мерах, схожи с предыдущей группой за исключением того, что на выходе ансамбля используется мера, объединяющая ранги от различных алгоритмов в единую метрику, на основе которой и делается вывод о включении или не включении признака в финальных набор данных.

Рассмотрены алгоритмы, предложенные различными авторами, оценено, как одни работы оказали влияние на другие и в целом на область отбора признаков.

Выводы. Результаты исследования могут быть применимы в различных областях. Например, исследователи могут получить исчерпывающую информацию об уже разработанных алгоритмах, что позволит им сократить время на теоретический обзор существующих подходов. Для прикладного применения данное исследование может быть отправной точкой в тестировании гипотез о том, какой алгоритм отбора признаков окажется наилучшим в конкретной текущей задаче.

Глухов В.Н. (автор) Подпись

Сметанников И.Б. (научный руководитель) Подпись