УДК 519.226.3

ОБУЧЕНИЕ БОЛЬШИХ БАЙЕСОВСКИХ СЕТЕЙ НА ОСНОВЕ ЭКСПЕРТНЫХ ЗНАНИЙ О БЛИЗОСТИ И КЛАСТЕРИЗАЦИИ ХАРАКТЕРИСТИК

Бубнова А.В. (Университет ИТМО)
Научный руководитель – к.т.н, доцент Калюжная А.В.
(Университет ИТМО)

Аннотация
В данной работе оценивается влияние на скорость и качество обучения больших байесовских сетей включения знаний о близости признаков, полученных с помощью кластеризации информационных расстояний или экспертных знаний. Реализован метод кластеризации, основанный на нормализованной взаимной информации. Локальные структуры, включаются в структуру большой байесовской сети в виде блоков, связанных через шифрование комбинаций значений.

Введение.
Существует огромное количество методов обучения структуры байесовской сети на данных. Одно из основных направлений обучения можно определить как задачу оптимизации выбранной целевой функции в пространстве ориентированных графов без циклов (DAG). Общей проблемой для этого направления является то, что само пространство поиска начинает расти суперэкспоненциально по мере роста числа признаков объекта. Скорость алгоритмов критически падает, а проблема выбора локально оптимальных структур становится еще более острой.
Растущий объем данных и практическая необходимость в многомерных моделях мотивирует к поиску быстрых алгоритмов обучения. В настоящее время можно выделить два направления. Первое заключается в распараллеливании поиска, однако не все целевые функции, особенно классические, такие как K2 и взаимная информация (MI), могут быть распараллелены. Второй подход опирается на методы машинного обучения, такие как kNN и кластеризации, для извлечения локальных структур и сжатия пространства поиска. Однако это порождает проблемы со включением подсетей в общую структуру без возникновения коллизий и возможной потерей качества.
Еще одна трудность при таком подходе — определение "расстояния" между двумя различными характеристиками. Обычные метрики не подходят, поскольку признаки могут быть разной размерности и природы. Например, пол, который является категориальной переменной, и количество лет. Существующие подходы, как правило, имеют дело с одними и теми же типами данных и опираются на различные информационные дивергенции, которые применяются для сравнения распределений. Такие дивергенции часто предполагают работу с дискретными распределениями, в то время как данные могут иметь смешанную природу. Решением является дискретизация непрерывных данных, но очевидно, что это может привести к потере информации.
Включение подсетей в общую сеть в случае kNN приводит к большому количеству циклов, и требуются дополнительные методы для приведения такой структуры к DAG. В случае кластеризации дополнительно необходимо решить вопрос о соединении кластеров в общую структуру. В данной работе эта проблема решается путем включения псевдопеременных, содержащих сжатую информацию о наиболее распространенных комбинациях значений в кластере. Существуют подходы, основанные на этой идее, но реализованные в контексте иерархических байесовских сетей, предполагающих древовидную структуру графа. В данной реализации, однако, нет никаких ограничений на структуру.

Основная часть.

Построение сети начинается с предварительной обработки данных. Все категориальные значения заменяются номерами категорий, а непрерывные — номерами бинов при дискретизации. Экспертные знания позволяют без затрат времени составлять кластеры, т. е. списки связанных переменных, но они не всегда доступны и статистически обоснованы. Целесообразно начать с расчета информационных расстояний и агломеративной кластеризации, которые могут предоставить такие списки без участия эксперта и подтвердить или опровергнуть экспертные знания. На этом этапе результаты кластеризации можно контролировать, задавая количество кластеров или пороговый уровень (treshold). Это позволяет найти баланс, поскольку при слишком маленьких кластерах пространство поиска сокращается незначительно, а при слишком больших увеличивается время поиска локальных структур, а также существует риск снижения качества, т. к. некоторые комбинации значений могут редко встречаться в данных.

На следующем этапе для каждого кластера генерируется псевдопеременная с зашифрованными значениями. Сначала выбираются наиболее представленные комбинации, например, охватывающие 95% всех вариантов, и им присваиваются номера. Для оставшихся вариантов рассматривается сходство с выбранными, в данной реализации с использованием расстояния Хэмминга, и затем им присваивается номер самой близкой комбинации.

Для каждого кластера производится поиск локальной структуры классическими методами. Отдельно обучается структура на зашифрованных переменных и переменных, не входящих ни в один кластер. Затем каждая псевдопеременная раздваивается на "входящую" и "исходящую" для формирования общей сети. Входящие и исходящие ребра от исходной переменной разделяются между ними. Между этими вершинами включается локальная структура соответствующего кластера. Дополнительные ребра проводятся от "входящих" ко всем вершинам кластера и от них к "исходящим". Выполняется параметрическое обучение всей структуры, при котором оцениваются условные вероятности для каждой пары вершин и всех вершин, из которых к ней ведёт ребро.

Критериями качества являются скорость работы алгоритма, а также точность в задаче восстановления значений, измеряемая в доле правильных категорий, предложенных моделью (accuracy). Алгоритм был протестирован на 10 подвыборках социальных наборов данных размера 1500, 3000, 4500, 6000, содержащих информацию о 50 различных характеристиках. А также на наборе геологических данных о 500 месторождениях нефти и газа, содержащих 15 переменных категориального и непрерывного типов.

Сравнение результатов показывает, что внедрение кластеризаций или экспертных знаний сокращает время обучение структур пропорционально количеству переменных и объему данных и дает сопоставимые результаты в задаче восстановления значений. А также улучшается интерпретируемость структуры сети.

Выводы.

Эта реализация может стать основой для быстрых алгоритмов обучения байесовских сетей, учитывающих имеющиеся знания или локальные особенности данных. Планируется развитие в сторону новых путей адаптации непрерывных данных, таких как включение информации о кластеризации в шифры, что потенциально повысит точность. А также более детальный анализ влияния гиперпараметров и параметров набора данных, что позволит определить ситуации, в которых может быть применен данный подход.