УДК 004.942

МЕТОД ОЦЕНКИ ОДНОРОДНОСТИ ВЫБОРОК АНАЛОГОВ С ИСПОЛЬЗОВАНИЕМ БАЙЕСОВСКИХ СЕТЕЙ

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

Аннотация
В данной работе оценивается однородность поднаборов, сгруппированных вокруг фиксированных элементов данных. Реализован метод оценки и сравнения существующих способов поиска таких подмножеств, который в качестве критерия использует проверку на унимодальность. Метод основан на вероятностном представлении подвыборок с помощью байесовских сетей, позволяющем корректно оценить необходимые свойства распределения.

Введение.

Как правило, методы, лежащие в основе вероятностного анализа, предполагают работу с выборкой, все элементы которой взяты из одной и той же генеральной совокупности или иными словами, что эти данные однородны или гомогенны. Наличие объектов из какого-то другого распределения сказывается на точности оценок, ухудшает качество моделей. На практике это может вызвать потери при неверном планировании и оценке рисков, затраты на получение дополнительной информации, проясняющей ситуацию. Возникает вспомогательная задача оценки однородности и поиска однородных подвыборок из объектов с десятками атрибутов, часть из которых связана сложной системой зависимостей. На решение данной проблемы и направлена эта работа.

Важным в контексте нашей задачи также является определение неоднородности. Под этим подразумевается наличие элементов хотя бы из двух различных генеральных совокупностей. Для оценки требуются методы описания генеральной совокупности по набору многомерных объектов, которые потенциально позволяют сравнивать соответствующие распределения и учитывают особенности конкретного набора данных.

Рассмотрим существующие подходы к решению задачи оценки и поиска однородных подвыборок и выявлению неоднородностей. В-первую очередь это различные методы кластеризации. Они распространены, существует множество удобных инструментов, однако также есть ряд проблем. Среди них: адаптированность части алгоритмов только под определенные типы данных, NP-полнота некоторых задач. Во-вторых, существуют тесты на однородность, но большинство из них корректно работает только при выполнении определенных условий, а также не дает информации о том, как искать подвыборки для проверки.

Также следует упомянуть метод поиска аналогов, который вычленяет подвыборку близкую к некоторому зафиксированному объекту. Потенциально он позволяет найти более однородное подмножество без полного разбиения данных и проверки неоднородности. К сожалению, нельзя гарантировать, что аналоги взяты из одной генеральной совокупности, т. к. не всегда удается найти элементы схожие по всем переменным одновременно и, следовательно, требуется дополнительная оценка однородности. Данная работа и посвящена решению этой редуцированной задачи.

Основная часть.

Поиск аналогов начинается с представления объектов в виде векторов в многомерном пространстве. Для решения задачи поиска ближайших можно использовать любые расстояния, работающие с соответствующими типами переменных. Поскольку атрибуты объектов могут быть как категориальными, так и количественными, далее используется

косинусное расстояние и коэффициент сходства Гауэра со стандартной нормализацией. Однако, это не единственный доступный вариант и, в общем случае, для нашего подхода к оценке не имеет значения каким способом подбирается набор похожих элементов.

Для представления аналогов в виде некоторого распределения используем дискретные байесовские сети (БС). Это направленный ацикличный граф (DAG), который хранит информацию о структуре зависимостей и условных распределениях. Заметим, что любая генеральная совокупность представима в виде некоторой байесовской сети. Из преимуществ стоит упомянуть, что для проверки на неоднородность можно использовать различия на уровне структуры графа или условных вероятностей. БС также позволяют оценить важные свойства распределений, например моды, пригодные для косвенной оценки однородности.

Граф и распределения можно использовать, но здесь возникает проблема разбиения аналогов на подвыборки для сравнения. Сама подзадача является вычислительно сложной и требует дополнительного исследования. По этой причине наш метод будет опираться на оценку мод, извлеченную из сети на подмножестве аналогов. В данном контексте моды интересны тем, что и для категориальных, и для количественных переменных наличие двух часто встречающихся значений, свидетельствует о том, что подвыборка разбивается на две части, отличные по распределению. Специфика аналогов заключена в том, что часть значений должна быть сконцентрирована вокруг соответствующего для целевого объекта. Тогда несовпадение с ним моды можно рассматривать как нарушение однородности, а частоту таких событий (accuracy) или среднюю величину отклонения (MAE) как количественную меру однородности аналогов, полученных с помощью определенного расстояния.

Для реализации этой идеи взяли набор данных из 500 многомерных объектов, содержащих информацию о 5 категориальных и 5 количественных атрибутах нефтегазовых месторождений. Для каждого объекта рассматривали 40 ближайших объектов-аналогов либо весь набор, обучали на них байесовские сети и с помощью семплирования проверяли совпадает ли мода или насколько отклоняется от значения для категориальных и количественных, соответственно. Надо заметить, что для построения дискретных байесовских сетей потребовалась предварительная дискретизация количественных переменных с помощью квантильных интервалов. На базе этого были рассчитаны частоты и среднее отклонение для каждой переменной, которые рассматриваются как количественные оценки однородности для каждого подхода. Сравнение результатов показало повышение однородности на категориальных переменных для всех рассмотренных расстояний относительно полного набора. Для количественных, однако, нельзя сделать такой же однозначный вывод.

Выводы.

Данная оценка является первым, но важным шагом при проверке однородности. Однако, ясно, что унимодальность не гарантирует однородность и необходимо расширить данный метод за счет более подробного исследования структуры и условных распределений. Учитывая неопределенные результаты на количественных переменных в будущем следует повторить этот эксперимент на более совершенных смешанных сетях.

Учитывая, что подбор аналогов часто используется при планировании разработки новых месторождений, данный метод может быть использован для поиска оптимального расстояния с точки зрения однородности аналогов.