АЛГОРИТМЫКЛАСТЕРИЗАЦИИДАННЫХ, ПОЛУЧЕННЫХИЗОТКРЫТЫХИСТОЧНИКОВСЕТИИНТЕРНЕТ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Крошнер А. О.

Самаль Д. А. – доцент, канд. техн. наук

В настоящее время в связи с развитием информационных технологий нарастает необходимость классификации и кластеризации данных, отмечается необходимость учитывать все большее количество параметров и характеристик при анализе объектов или явлений. Как итог - невероятный спрос на разработку и применение методов для классификации и кластеризации одномерных и многомерных данных, поиска закономерностей.

Кластеризация представляет из себя задачу по разбиению множества объектов на кластеры. Отличительными признаками кластера являются однородность и изолированность. Однородность означает, что внутри кластера будут находиться схожие между собой объекты. Изолированность же показывает, что объекты одного кластера имеют как можно больше отличий от объектов в другом кластере. В кластеризации, в противовес классификации, перечень классов изначально не задан и определяется по ходу выполнения алгоритма.

Основной проблемой кластеризации данных является отсутствие универсального решения и, как следствие, необходимость выбора алгоритмов и их параметров экспериментальным путем. Выделение универсального решения задачи кластеризации невозможно по следующим причинам:

- Разнородность данных и необходимость их предварительного преобразования.
- Выбор параметров объекта, которые будут использоваться в процессе классификации.
- Определение меры расстояния (сходства/подобия) между объектами.
- Выбор метода кластеризации.
- Заранее неизвестное итоговое количество кластеров.

Алгоритмы кластеризации могут быть классифицированы по результату работы алгоритма. Выделяют неиерархические плоские алгоритмы (на выходе - некоторое множество кластеров) и иерархические древовидные (на выходе - дерево кластеров с некоторой степенью вложенности).

Плоские неиерархические алгоритмы в свою очередь подразделяются на четкие и нечеткие. В четких алгоритмах кластеризуемый объект может принадлежать исключительно одному кластеру, в нечетких алгоритмах вычисляется вероятность принадлежности объекта к полученным кластерам.

Иерархические алгоритмы кластеризации подразделяются на агломеративные и дивизивные. В агломеративных алгоритмах на начальной итерации каждому объекту ставится в соответствие кластер, которые объединяются в более крупные кластеры при последующих итерациях алгоритма. Дивизивные алгоритмы представляют из себя противоположность агломеративным: на начальной итерации все объекты состоят в одном кластере, а при последующих итерациях кластеры дробятся на более мелкие. Графическая визуализация результатов работы иерархических алгоритмов может быть представлена в виде дендрограммы, которая может быть построена с использованием матрицы мер близости. Дендрограмма позволяет отобразить взаимные связи объектов для исходного множества.

Основными метриками, которые могут быть использованы в качестве меры близости объектов, являются: Эвклидово расстояние, мера сходства Хэмминга, мера сходства Роджерса-Танимото, Манхэттенская метрика, Расстояние Махаланобиса, Расстояние Журавлева.

Зачастую исследователи, занимающиеся вопросами получения некого знания из данных (data mining и data clustering как частный случай), сталкиваются с вопросами при постановке задачи (классический data mining, получение нового знания из набора данных), выборе способа решения задачи и интерпретации полученных результатов. Результаты исследования могут не согласоваться с мнением и опытом других исследователей и профессионалов в исследуемом домене.

Проблемы кластерного анализа вытекают из природы объекта исследования. Выделяются следующие критерии (требования), предъявляемые к исследуемым данным

1. Корреляция данных между собой: зачастую в кластерном анализе в качестве меры сходства данных используется Евклидова расстояние и различные варианты, поэтому описательные шкалы должны быть ортонормированы, т.к. в противном случае применение Евклидовой меры теоретически некорректно и необоснованно. Результатом этого критерия являются проблемы выбора метрики для не ортонормированных пространств и ортонормирование пространства.
2. Безразмерность показателей (характеристик исследуемых данных): исследователь, выбирая единицу измерения характеристики, существенно влияет на результаты исследования. Выбор зачастую является произвольным, соответственно и результаты могут оказаться не объективными. Результатом этого критерия является проблема сопоставимости характеристик, исследуемых данных, а также их формализация.

3. Показатели должны быть распределены равномерно: требование происходит из того, что обоснование корректности перечисленных выше метрик (мер близости объектов) основано на нормальности распределения этих объектов. Результатом этого критерия являются проблемы необходимости доказательства нормального распределения объектов исходной выборки и/или их нормализация.
4. Устойчивость данных (выборки) к влиянию случайных факторов
5. Однородность данных (выборки): 4-ое и 5-ое требования взаимосвязаны между собой, случайные факторы как-раз и приводят к выбросам и шуму. Результатом этого критерия является необходимость для исследователя учитывать неоднородность данных и зашумление, принимать во внимание возможность детектирования выбросов и фильтрации и принятия решения о маркировке таких данных: первоначальное гипотеза о степени зашумленности исходных данных может быть ошибочной. Вследствие этого может быть полезно отобразить промаркированные выбросы в результирующем наборе особым образом.

Следует отметить, что перечисленные требования на практике в полной мере выполняются далеко не всегда, а иногда некоторые из них сознательно игнорируются, т.к. предварительная обработка (processing) данных может довольно сильно повлиять на размерность данных, переданных для исследования.

Работа, проделанная мной в рамках дипломного проекта по алгоритмам коллаборативной фильтрации пользователей социальных сетей, показала, что доступность данных и их однородность имеет огромное значение. Под доступностью подразумевается отсутствии необходимости копировать данные из удаленного локально небольшими порциями вследствие ограничений API социальных сетей. Еще одним минусом данных, которые исследовать может попробовать получить через API сторонних программных продуктов является их неоднородность и неактуальность из-за отсутствия серьезных ограничений на формат данных, получаемых такими программными продуктами (пример – минимальное количество обязательных для заполнения полей и, как результат, большое количество пустых данных в загружаемой выборке). Решением видится использование больших, открытых наборов данных (датасетов), предлагаемых для загрузки на локальный компьютер/вычислительный кластер.

Среди стран СНГ Республика Беларусь довольно сильно отстает по возможности получения данных от соседей, таких как Украина и Российская Федерация. Тем не менее, на ресурсе [4] можно найти довольно интересные выборки. Стоит отметить, что большая часть выборок представляет из себя агрегированную статистику с малым количеством полей, в отличии от данных, которые доступны благодаря США [5] или ЕС с большим количеством данных по актуальным демографическим, технологическим и климатическим вопросам.

Анализ датасетов в сети Интернет показал их достаточное количестве и разнообразие, вследствие этого исследователю, желающему проделать какую-либо работу над данными следует выбрать такой домен, в котором исследователь имеет хотя бы небольшую экспертизу, или может привлечь сторонних экспертов как для постановки задачи, так и для обработки результатов анализа.

Многие социальные сети выставляют наружу публичный API, который дает доступ к внутренней базе данных. Примеры методов API: пользователи (информация по каждому из пользователей с публичным профилем), связи междупользователи внутри приложения, публикации, отмеченные пользователем. API сервисов активно пользуются как злоумышленники, так и другие компании, которые, применив собственные реализации и комбинации алгоритмов анализа данных пользователей, могут извлечь некоторое полезное «знание», которое может быть использовано по их усмотрению. Одним из таких сервисов является Git.FM – некоммерческий сервис, который предлагает рекомендации для пользователей проекта GitHub – площадка для хостинга исходного кода приложений, одна из наиболее популярных платформ, которой пользуется большое количество продуктов с открытым исходным кодом. Git.FM предлагает своему пользователю рекомендации (на основе активности пользователю на GitHub), какому проекту можно уделить свое время. Таким образом, пользователь, доверившись результатам работы сервиса Git.FM, больше не должен самостоятельно искать себе проекты, сервис выполняет эту работу за него.

Еще одним примером сервиса, который использует в своей работе алгоритмы кластеризации и рекомендации является RetailRocket. Сервис предоставляет интернет-магазинам возможность добавить систему рекомендаций на свой сайт. API данного программного продукта предлагает связать пользовательскую корзину покупок (действия с ней, такие как добавить товар, удалить товар) со своим продуктом, построить кластеры схожих пользователей и использовать результаты для дальнейшей обработки (процессинга – создания набора рекомендованных пользователю товаров). Таким образом, для интернет-площадок отпадает необходимость в реализации рекомендательной системы самостоятельно.

Список использованных источников:

1. Мандель И. Д. Кластерный анализ. М.: Финансы и статистика, 1988. – 176с.
2. Sammut C., Webb J. Encyclopedia of Machine Learning – NY, USA, 2010.
3. Луценко Е.В., Коржаков В.Е. Некоторые проблемы классического кластерного анализа – журнал Адыгейского Государственного Университета, выпуск 2, 2011.
4. Портал «Открытые данные Беларуси» - https://opendata.by/
5. Портал «Открытые данные США» - https://www.data.gov/