Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей
Кафедра информатики

РЕФЕРАТ

на тему

Кластерный анализ и его задачи. Итерационные методы. Иерархические методы (меры сходства)

Магистрант: А.С. Долматович

МИНСК 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Главная цель кластерного анализа — нахождение групп схожих объектов в выборке данных. Эти группы удобно называть кластерами. Общепринятого определения понятия «кластер» не существует, однако очевидно, что кластер может быть охарактеризован рядом признаков, «...наиболее важными из которых являются плотность, дисперсия, размеры, форма и отделимость»

Б.Г. Миркин отмечает, что «самое естественное — это дать строгое определение компактной группы, после чего конструировать группировку как совокупность компактных в смысле данного определения групп». Так, он предлагает называть группу объектов кластером, если максимальное расстояние между ее точками не превышает минимального расстояния «во вне», т.е. минимального расстояния между точками, попавшими в группу и не попавшими в нее. Возможно использование и более слабого требования к компактности группы: сгущением можно назвать такое множество точек, для которого среднее внутреннее расстояние меньше, чем среднее расстояние во вне.

ИСТОРИЯ

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом. Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству.

Термин кластерный анализ в действительности включает в себя набор различных алгоритмов кластеризации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры, т.е. развернуть таксономии. Например, биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии, человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. Стоит заметить, что в этой классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Человек имеет больше сходства с другими приматами (т.е. с обезьянами), чем с "отдаленными" членами семейства млекопитающих (например, собаками) и т.д.

ПРЕДЛОЖЕНИЕ

Задача кластеризации заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m (m – целое) кластеров (подмножеств) Q1, Q2, …, Qm, так, чтобы каждый объект Gj принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2$$

где xj - представляет собой измерения j-го объекта.

Для решения задачи кластерного анализа необходимо определить понятие сходства и разнородности.

Понятно то, что объекты i-ый и j-ый попадали бы в один кластер, когда расстояние (отдаленность) между точками Xi и Xj было бы достаточно маленьким и попадали бы в разные кластеры, когда это расстояние было бы достаточно большим. Таким образом, попадание в один или разные кластеры объектов определяется понятием расстояния между Xi и Xj из Ep, где Ep - p-мерное евклидово пространство.

Наиболее часто употребляются следующие функции расстояний:

1. Евклидово расстояние

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p} \left(x_{ki} - x_{kj}\right)^2\right]^{\frac{1}{2}}$$

2. l1 - норма

$$d_1(X_i, X_j) = \left[\sum_{k=1}^{p} \left| x_{ki} - x_{kj} \right|\right]$$

3. Супремум - норма

$$d_{\infty}(X_i, X_j) = \sup_{k = 1, 2, \ldots, p} \left\{ \left| x_{ki} - x_{kj} \right| \right\}$$

$$d_p(X_i, X_j) = \left[\sum_{k=1}^{p} \left| x_{ki} - x_{kj} \right|^p\right]^{1/p}$$

4. lp - норма

Методы кластеризации

Сегодня существует достаточно много методов кластерного анализа. Остановимся на некоторых из них (ниже приводимые методы принято называть методами минимальной дисперсии).

Пусть X - матрица наблюдений: X = (X1, X2,..., Xu) и квадрат евклидова расстояния между Xi и Xj определяется по формуле:

$$d_{ij}^2 = \left(X_i - X_j\right)^T \left(X_i - X_j\right)$$

1) Метод полных связей.

Суть данного метода в том, что два объекта, принадлежащих одной и той же группе (кластеру), имеют коэффициент сходства, который меньше

некоторого порогового значения S. В терминах евклидова расстояния d это означает, что расстояние между двумя точками (объектами) кластера не должно превышать некоторого порогового значения h. Таким образом, h определяет максимально допустимый диаметр подмножества, образующего кластер.

2) Метод максимального локального расстояния.

Каждый объект рассматривается как одноточечный кластер. Объекты группируются по следующему правилу: два кластера объединяются, если максимальное расстояние между точками одного кластера и точками другого минимально. Процедура состоит из n - 1 шагов и результатом являются разбиения, которые совпадают со всевозможными разбиениями в предыдущем методе для любых пороговых значений.

3) Метод Ворда.

В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

4) Центроидный метод.

Расстояние между двумя кластерами определяется как евклидово расстояние между центрами (средними) этих кластеров:

$$d2\ ij = (\grave{}X - \grave{}Y)T(\grave{}X - \grave{}Y)$$

Кластеризация идет поэтапно на каждом из n–1 шагов объединяют два кластера G и p, имеющие минимальное значение d2ij Если n1 много больше n2, то центры объединения двух кластеров близки друг к другу и характеристики второго кластера при объединении кластеров практически

игнорируются. Иногда этот метод иногда называют еще методом взвешенных групп.

Иерархические методы порождают все возможные варианты построения кластеров при выбранной эвристической мере сходства.

Вариант построения всевозможных кластеров обычно изображают в виде дерева, которое показывает схему объединения объектов в кластеры. Каждому уровню объединения соответствует максимальное значение целевой функции, достигаемое на каждом уровне группировки объектов. В качестве целевой функции используется мера сходства, вычисленная на всей совокупности разбиваемых объектов и выделяемых кластеров.

Следует также обратить внимание, что при количестве объектов, превышающих десятки, создаваемое дерево будет очень большим. Безусловно, невозможно изучить такое огромное число вариантов построения кластеров. Обилие вариантов, порождаемых иерархическими процедурами, затрудняет их анализ. Поэтому метод дендограмм находит применение при построении гипотез о существовании сегментов на малых выборках, например на данных, полученных при проведении фокус-групп, которые насчитывают несколько десятков респондентов. Практически результаты анализа дендограмм могут быть использованы для построения гипотез о числе выделяемых кластеров для больших выборок (сотен и тысяч респондентов), которые обрабатываются Kmeans-алгоритмами.

Большей популярностью, чем иерархические подходы, при сегментировании пользуются итерационные методы. Это объясняется тем, что на практике необходимо выделять сегменты на больших выборках.

Эти методы определяются как пошаговое подключение к одному из формируемых кластеров одного объекта, который выбирается по одной из возможных итерационных схем кластеризации. Кроме того, по способу выбора первых объектов для каждого из создаваемых кластеров итерационные методы разделяются два типа: автоматические и экспертные.

Автоматические методы (Kmeans-алгоритмы)

Большинство итерационных методов выполняется следующим образом. В качестве исходного разбиения принимается гипотеза о существовании конкретного количества кластеров, которые необходимо выделить.

На первом шаге указывается это конечное число кластеров (гипотеза о К-кластерах).

На втором шаге для каждого из кластеров вычисляется центр кластера. Например, в качестве центра может быть выбран произвольный объект. В практических реализациях центр кластера вычисляется на основе случайных выборок, которые выделяются из всей совокупности анализируемых потребителей.

После определения центров кластеров просматриваются объекты из анализируемой совокупности потребителей. Согласно установленной мере сходства выбирается тот объект, который имеет лучшую меру сходства по отношению к другим объектам и формируемым кластерам. По выбранной мере объект относится к одному из кластеров.

Далее процесс вычисления продолжается до тех пор, пока все объекты не будут разнесены по кластерам.

Экспертная кластеризация

При экспертном сегментировании аналитику предоставляется возможность самому указать, какие из объектов целесообразно включить в кластер как образцы. Обычно, исходя из своего понимания потребителей как объектов кластеризации, аналитик может достаточно точно сказать, какие из объектов следует отличать друг от друга. Конечно, такое предположение аналитик делает только на интуитивном уровне. Он не может оценить весь комплекс переменных, которыми описываются объекты. Данная процедура позволяет учесть интуитивные знания аналитика о принадлежности клиентов с различными свойствами к исследуемым сегментам.

С учетом установленной принадлежности к сегментам отдельных представительных образцов вычисляются центры формируемых кластеров. А

затем для определенных экспертным образом центров формируются кластеры путем изучения всей совокупности потребителей.

Существуют различные модификации этой схемы, однако практик-аналитик может опустить эти тонкости реализаций. Экспертный подход к выделению сегментов можно рассматривать как альтернативу методу дендограмм.

Генетические алгоритмы кластеризации

Можно попытаться улучшить найденный вариант разбиения путем перемещения объектов из одного кластера в другой. Для этого применяются специальные алгоритмы, смысл которых сводится к улучшению общей целевой функции, построенной на мерах сходства, через выбор по определенному критерию объектов, перемещаемых между кластерами. Эта идея, в частности, реализуется в генетических алгоритмах.

Эти алгоритмы выполняют перемещения объектов между кластерами, и на каждой итерации перераспределения объектов между кластерами учитываются наиболее продуктивные комбинации размещения, построенные на предшествующих шагах поиска.

Можно сказать, что традиционный эвристический алгоритм позволяет найти «пробный» вариант разбиения потребителей по кластерам, а генетический алгоритм пытается его улучшить.

Заметим, что кроме генетических алгоритмов для кластеризации в последние годы широко применяются нейронные сети. Однако их применение в практических исследованиях сегментов требует от аналитиков дополнительной математической подготовки.

НЕДОСТАТКИ

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения: В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

В кластерном анализе считается, что:

а) выбранные характеристики допускают в принципе желательное разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Выбор масштаба играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклоненение, так что дисперсия оказывается равной единице.

ДОСТОИНСТВА

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы социально-экономической информации, делать их компактными и наглядными.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие (например, общехозяйственной и товарной конъюнктуры). Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл здесь может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

ЗАКЛЮЧЕНИЕ

Изложенное выше позволяет сделать вывод о том, что применению метода кластерного анализа должно предшествовать изучение теории и накопленной практики этого использования. На начальных этапах использования этого метода исследователь должен иметь четко представление, которое из двух задач он решает. Или это обычная задача типизации, при котором исследуемую совокупность наблюдений следует разделить на относительно небольшое количество группировок. Тогда выполняется работа, аналогичная получению интервалов статистического группировки при обработке одномерных наблюдений. При этом операция осуществляется так, чтобы элементы одной области группировки находились друг от друга по возможности на расстоянии. Вторая задача может заключаться в том, что исследователь пытается определить естественную расстояние выходных элементов (наблюдений) на четко выраженные кластеры, находящиеся друг от друга на некотором расстоянии, но которые не разбиваются на такие же отдаленные друг от друга части. Следует помнить, что первая задача (задача типизации) всегда имеет решение, второе - в своей постановке может иметь отрицательный результат, то есть может оказаться, что множество выходных наблюдений не проявляет природного расположения на кластеры, например, образует один кластер.

Немаловажным этапом кластерного анализа является выбор переменных (признаков). Эта стадия анализа является основой формирования одинаковых пространств, в которых должно проводиться моделирование.

Выбор признаков осуществляется, как правило, в две стадии. В основе первой из них лежит формирование первичной гипотезы о наборе признаков, влияющих на изучаемое явление; в основе второй - уточнение гипотезы по результатам консультаций (опросов) специалистов исследуемой области.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

HABR [Электронный ресурс]. - Электронные данные. - Режим доступа: https://habr.com/ru/company/ods/blog/328372/ - Дата доступа: 04.11.2019.

HABR [Электронный ресурс]. - Электронные данные. - Режим доступа: https://habr.com/ru/company/jetinfosystems/blog/420261/ - Дата доступа: 05.11.2019.

WIKIPEDIA [Электронный ресурс]. - Электронные данные. - Режим доступа: https://en.wikipedia.org/wiki - Дата доступа: 07.11.2019.

IFMO [Электронный ресурс]. - Электронные данные. - Режим доступа: http://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче_кластеризации - Дата доступа: 12.11.2019.