Исследование алгоритмов кластерного анализа для решения задач WEB Mining на основе данных, полученных билетной платформой BIL24

ВВЕДЕНИЕ

Кластерный анализ применяется в различных областях. Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Для решения экономических задач кластеризация длительное время мало использовалась из-за специфики экономических данных и явлений. Ярким примером задачи кластеризации в маркетинге является разбиение персонала на различные группы, например, по психологической совместимости, классификация потребителей и поставщиков поцелевому признаку, выявление схожих производственных ситуаций, при которых возникает брак. В социологии задача кластеризации – разбиение респондентов на однородные группы. В маркетинговых исследованиях кластерный анализ применяется достаточно широко – как в теоретических исследованиях, так и практикующими маркетологами, решающими проблемы группировки различных объектов. При этом решаются вопросы о группах клиентов, продуктов и т.д. Одной из наиболее важных задач при применении кластерного анализа в маркетинговых исследованиях является анализ поведения потребителя, а именно: группировка потребителей в однородные классы для получения максимально полного представления о поведении клиента из каждой группы и о факторах, влияющих на его поведение. Важной задачей, которую может решить кластерный анализ, является позиционирование, т.е. определение ниши, в которой следует позиционировать новый продукт, предлагаемый на рынке. В результате применения кластерного анализа строится карта, по которой можно определить уровень конкуренции в различных сегментах рынка и соответствующие характеристики товара для возможности попадания в этот сегмент. С помощью анализа такой карты возможно определение новых, незанятых ниш на рынке, в которых можно предлагать существующие товары или разрабатывать новые.

Кластерный анализ также может быть удобен, например, для анализа клиентов различных компаний. Для этого все клиенты группируются в кластеры, и для каждого кластера вырабатывается индивидуальная политика. Такой подход позволяет существенно сократить объекты анализа, и, в то же время, индивидуально подойти к каждой группе клиентов.

Таким образом, кластерный анализ занимает одну из лидирующих позиций в интеллектуальном анализе данных. Он помогает упростить работу с большим объёмом информации, компактно визуализировать данные, выявлять новые закономерности и т.д.

Сегодня методы кластеризации активно используются в WEB для продвижения коммерческих продуктов, на основе выводов о пользовательской активности в Интернет.

Целью данной работы является изучение методов кластеризации, проведение экспериментов на основе данных, полученных билетной платформой BIL24, визуализация и анализ результатов. В дальнейшем планируется внедрение изученных методов и результатов их работы для маркетинговых исследований, билетной платформы BIL24.

1 Общие сведения о кластеризации

1.1 Понятие кластеризации

Кластеризация – это автоматическое разбиение элементов некоторого множества на группы в зависимости от их схожести. Синонимами термина «кластеризация» являются «автоматическая классификация», «обучение без учителя» и «таксономия». В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных – классы; по этим признакам новый объект можно отнести к тому или иному классу.

Задача кластеризации схожа с задачей классификации и является ее логическим продолжением, но её отличие в том, что классы изучаемого набора данных заранее не предопределены. Таким образом кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить, как точки в признаковом пространстве, то задача кластеризации сводится к определению «сгустков точек».

Кластеризация является описательной процедурой и не делает никаких статистических выводов, но дает возможность изучить структуру данных.

Само понятие «кластер» определено неоднозначно: в каждом исследовании свои «кластеры». Понятие «кластер» переводится как «скопление», «гроздь». В искусственных нейронных сетях под понятием кластер понимается подмножество «близких друг к другу» объектов из множества векторов характеристик. Следовательно, кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

1.2 Процесс кластеризации

Процесс кластеризации зависит от выбранного метода и почти всегда является итеративным. Он может включать множество экспериментов по выбору разнообразных параметров, например, меры расстояния, типа стандартизации переменных, количества кластеров и т.д. Однако эксперименты не должны быть самоцелью – ведь конечной целью кластеризации является получение содержательных сведений о структуре исследуемых данных. Полученные результаты требуют дальнейшей интерпретации, исследования и изучения свойств и характеристик объектов для возможности точного описания сформированных кластеров.

Кластеризация данных включает в себя следующие этапы:

а) Выделение характеристик.

Для начала необходимо выбрать свойства, которые характеризуют исследуемые объекты, ими могут быть количественные характеристики (координаты, интервалы), качественные характеристики (цвет, статус, воинское звание) и т.д. Затем стоит попробовать уменьшить размерность пространства характеристических векторов, то есть выделить наиболее важные свойства объектов. Уменьшение размерности ускоряет процесс кластеризации и в ряде случаев позволяет визуально оценивать результаты. Выделенные характеристики стоит нормализовать. Далее все объекты представляются в виде характеристических векторов. И объект полностью отождествляется с его характеристическим вектором.

б) Определение метрики.

Следующим этапом кластеризации является выбор метрики, по которой будет определяться «схожесть» объектов. Пусть d(хi, хj) – некоторая мера близости между каждой парой классифицируемых объектов i и j. В качестве таковой может использоваться любая полезная функция: евклидово или манхэттенское расстояние, коэффициент корреляции Пирсона, расстояние χ2, коэффициенты сходства Жаккара, Съеренсена, Ренконена и многие другие.

в) Представление результатов.

Результаты кластеризации должны быть представлены в удобном для обработки виде, чтобы осуществить оценку качества кластеризации. Обычно используется один из следующих способов:

- представление кластеров центроидами;

- представление кластеров набором характерных точек;

- представление кластеров их ограничениями.

Оценка качества кластеризации может быть проведена на основе следующих процедур:

- ручная проверка;

- установление контрольных точек и проверка на полученных кластерах;

- определение стабильности кластеризации путем добавления в модель новых переменных;

- создание и сравнение кластеров с использованием различных методов.

Разные методы кластеризации могут создавать разные кластеры, и это является нормальным явлением. Однако создание схожих кластеров различными методами указывает на правильность кластеризации.

1.3 Алгоритмы кластеризации

Следует отметить, что в результате применения различных методов кластерного анализа могут быть получены кластеры различной формы. Например, возможны кластеры "цепочного" типа, когда кластеры представлены длинными "цепочками", кластеры удлиненной формы и т.д., а некоторые методы могут создавать кластеры произвольной формы. Различные методы могут стремиться создавать кластеры определенных размеров (например, малых или крупных), либо предполагать в наборе данных наличие кластеров различного размера. Некоторые методы кластерного анализа особенно чувствительны к шумам или выбросам, другие - менее. В результате применения различных методов кластеризации могут быть получены неодинаковые результаты, это нормально и является особенностью работы того или иного алгоритма. Данные особенности следует учитывать при выборе метода кластеризации. На сегодняшний день разработано более сотни различных алгоритмов кластеризации.

Классифицировать алгоритмы можно следующим образом (рисунок 1):

1) Иерархические и плоские.

Иерархические алгоритмы (также называемые алгоритмами таксономии) строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений. Таким образом, на выходе получается дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластера.

Плоские алгоритмы строят одно разбиение объектов на кластеры.

2) Четкие и нечеткие.

Четкие (или непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит только одному кластеру. Нечеткие (или пересекающиеся) алгоритмы каждому объекту ставят в соответствие набор вещественных значений, показывающих степень отношения объекта к кластерам. Это значит, что каждый объект относится к каждому кластеру с некоторой вероятностью.

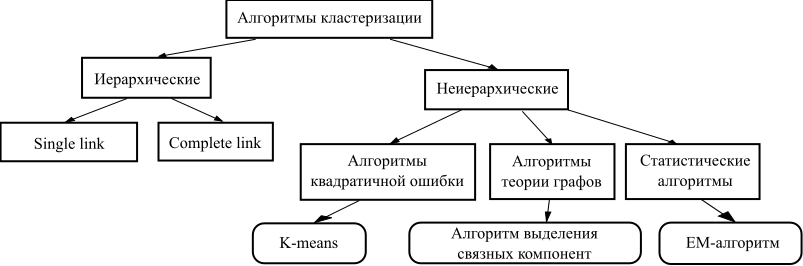


Рисунок 1 – Классификация алгоритмов кластеризации []

1.3.1 Иерархические алгоритмы

Результатом работы иерархических алгоритмов является дендограмма (иерархия), позволяющая разбить исходное множество объектов на любое число кластеров. Два наиболее популярных алгоритма, оба строят разбиение «снизу – вверх»:

- single – link (рисунок 2) – на каждом шаге объединяет два кластера с наименьшим расстоянием между двумя любыми представителями;

- complete – link (рисунок 3) – на каждом шаге объединяет два кластера с наименьшим расстоянием между двумя наиболее удаленными представителями.

Здесь укажите хотя бы что такое Х1 и Х2.

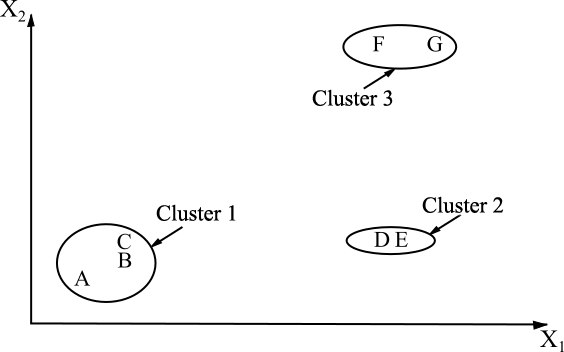


Рисунок 2 – Пример single – link алгоритма

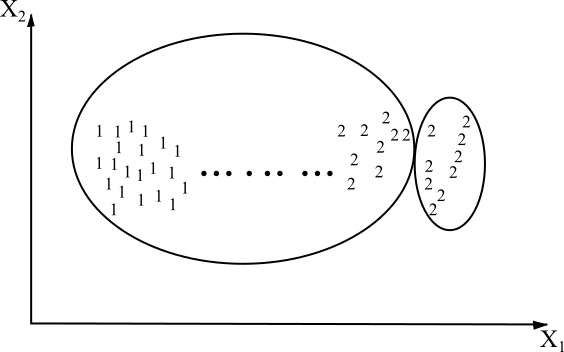


Рисунок 3 – Пример complete – link алгоритма

1.3.2 k-Means алгоритм

Данный алгоритм (рисунок 4) состоит из следующих шагов:

1. Случайно выбрать k точек, являющихся начальными координатами «центрами масс» кластеров (любые k из n объектов, или вообще k случайных точек).

2. Отнести каждый объект к кластеру с ближайшим «центром масс».

3. Пересчитать «центры масс» кластеров согласно текущему членству.

4. Если критерий остановки не удовлетворен, вернуться к шагу 2.

В качестве критерия остановки обычно выбирают один из двух: отсутствие перехода объектов из кластера в кластер на шаге 2 или минимальное изменение среднеквадратической ошибки.

Алгоритм чувствителен к начальному выбору «центр масс».

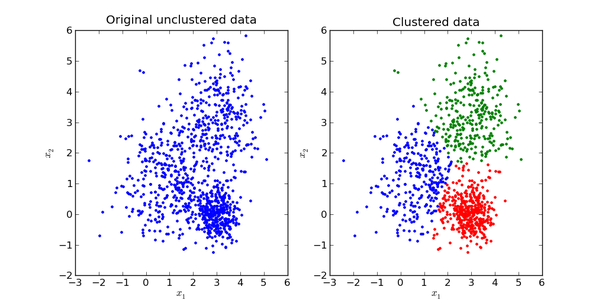


Рисунок 4 – Пример результатов работы k-Means алгоритма

1.3.3 Минимальное покрывающее дерево

Данный метод (рисунок 5) производит иерархическую кластеризацию «сверху вниз». Сначала все объекты помещаются в один кластер, затем на каждом шаге один из кластеров разбивается на два, так чтобы расстояние между ними было максимальным.

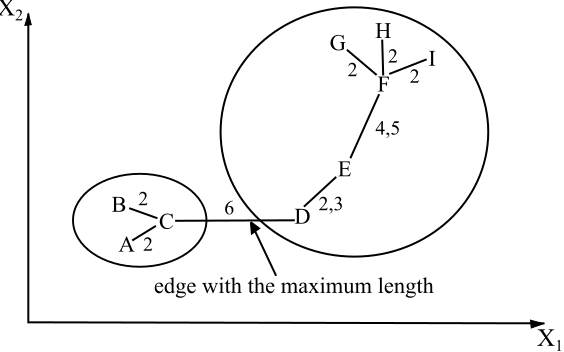


Рисунок 5 – Пример результата работы алгоритма минимального покрывающего дерева

1.3.4 Метод ближайшего соседа

Этот метод является одним из старейших методов кластеризации. Он был создан в 1978 году. Он прост и наименее оптимален из всех представленных.

Для каждого объекта вне кластера делаем следующее:

1. Находим его ближайшего соседа, кластер которого определен.

2. Если расстояние до этого соседа меньше порога, то относим его в тот же кластер. Иначе из рассматриваемого объекта создается еще один кластер.

Далее рассматривается результат и при необходимости увеличивается порог, например, если много кластеров из одного объекта.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1 Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Монография. — Тольятти; Лондон: Б.и., 2017. — 351 с.