ЛЕКЦИЯ 5 ЗАДАЧИ DATA MINING. КЛАССИФИКАЦИЯ И КЛАСТЕРИЗАЦИЯ

В предыдущей лекции мы кратко остановились на основных задачах *Data* Mining. Две из них - *классификацию* и *кластеризацию* - мы рассмотрим подробно в этой лекции.

Задача классификации

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей *Data* Mining. Ввиду распространенности задач *классификации* необходимо четкое понимание сути этого понятия.

Приведем несколько определений.

Классификация - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Различают:

- вспомогательную (искусственную) *классификацию*, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- естественную *классификацию*, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их *сочетания* и процедуры деления понятий *классификация* может быть:

- простой - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой *классификации* является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");

• сложной - применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое. Примером такой *классификации* является периодическая система химических элементов.

Под *классификацией* будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

Классификация - это *закономерность*, позволяющая делать *вывод* относительно определения характеристик конкретной группы. Таким образом, для проведения *классификации* должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или *объект* (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии *обучения с учителем* (*supervised learning*), которое также именуют контролируемым или управляемым обучением.

Задачей *классификации* часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Например, можно предсказать, кто из клиентов фирмы является потенциальным покупателем определенного товара, а кто - нет, кто воспользуется услугой фирмы, а кто - нет, и т.д. Этот тип задач относится к задачам ***бинарной классификации***, в них зависимая *переменная* может принимать только два значения (например, да или нет, 0 или 1).

Другой вариант *классификации* возникает, если зависимая *переменная* может принимать значения из некоторого *множества* предопределенных классов. Например, когда необходимо предсказать, какую марку автомобиля захочет купить клиент. В этих случаях рассматривается множество классов для зависимой переменной.

Классификация может быть ***одномерной*** (по одному признаку) и ***многомерной*** (по двум и более признакам).

Многомерная классификация была разработана биологами при решении проблем дискриминации для классифицирования организмов. Одной из первых *работ*, посвященных этому направлению, считают работу Р. Фишера (1930 г.), в которой организмы разделялись на подвиды в зависимости от результатов измерений их физических параметров. Биология была и остается наиболее востребованной и удобной средой для разработки *многомерных* методов *классификации*.

Рассмотрим задачу *классификации* на простом примере. Допустим, имеется *база данных* о клиентах туристического агентства с информацией о возрасте и доходе за месяц. Есть рекламный материал двух видов: более дорогой и комфортный отдых и более дешевый, молодежный отдых. Соответственно, определены два класса клиентов: *класс* 1 и *класс* 2. *База данных* приведена в таблице 5.1.

Таблица 5.1. База данных клиентов туристического агентства

Код клиента	Возраст	Доход	Класс
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2

Задача. Определить, к какому классу принадлежит новый клиент и какой из двух видов рекламных материалов ему стоит отсылать.

Для наглядности представим нашу базу данных в двухмерном измерении (возраст и доход), в виде *множества* объектов, принадлежащих классам 1 (оранжевая *метка*) и 2 (серая *метка*). На рис. 5.1 приведены объекты из двух классов.

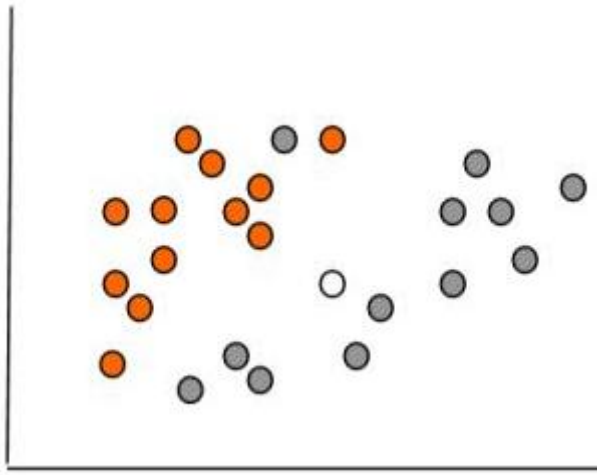

Рис. 5.1. Множество объектов базы данных в двухмерном измерениии

Решение нашей задачи будет состоять в том, чтобы определить, к какому классу относится новый клиент, на рисунке обозначенный белой меткой.

Процесс классификации

Цель процесса *классификации* состоит в том, чтобы построить модель, которая использует прогнозирующие атрибуты в качестве входных параметров и получает *значение* зависимого *атрибута*. *Процесс классификации* заключается в разбиении *множества* объектов на классы по определенному критерию.

Классификатором называется некая сущность, определяющая, какому из предопределенных классов принадлежит *объект* по вектору признаков.

Для проведения *классификации* с помощью математических методов необходимо иметь формальное описание объекта, которым можно оперировать, используя математический аппарат *классификации*. Таким описанием в нашем случае выступает *база данных*. Каждый *объект* (*запись базы данных*) несет информацию о некотором свойстве объекта.

Набор исходных данных (или выборку данных) разбивают на два *множества*: обучающее и тестовое.

Обучающее множество (*training set*) - множество, которое включает данные, использующиеся для обучения (конструирования) модели.

Такое множество содержит входные и выходные (целевые) значения примеров. Выходные значения предназначены для обучения модели.

Тестовое (test *set*) множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки работоспособности модели.

Процесс *классификации* состоит из двух этапов [21]: конструирования модели и ее использования.

1. Конструирование модели: описание множества предопределенных классов.
   - Каждый пример набора данных относится к одному предопределенному классу.
   - На этом этапе используется обучающее множество, на нем происходит конструирование модели.
   - Полученная модель представлена классификационными правилами, деревом решений или математической формулой.
2. Использование модели: *классификация* новых или неизвестных значений.
   - Оценка правильности (точности) модели.
   1. Известные значения из тестового примера сравниваются с результатами использования полученной модели.
   2. Уровень точности - процент правильно классифицированных примеров в тестовом множестве.
   3. Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.
   - Если точность модели допустима, возможно использование модели для *классификации* новых примеров, класс которых неизвестен.

Процесс *классификации*, а именно, *конструирование* модели и ее использование, представлен на рис. 5.2. - 5.3.

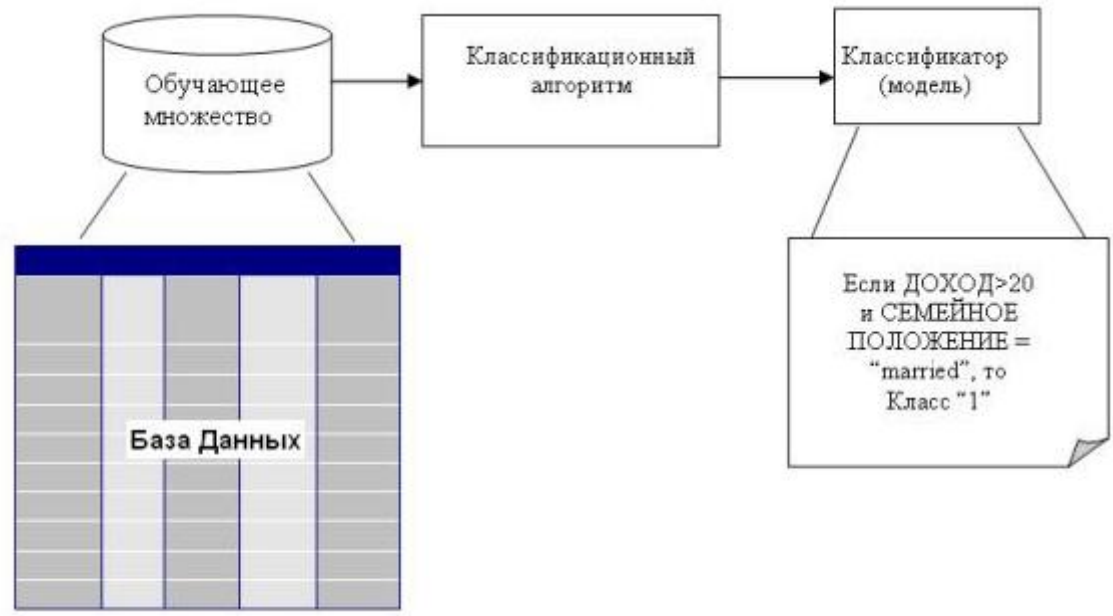

Рис. 5.2. Процесс классификации. Конструирование модели

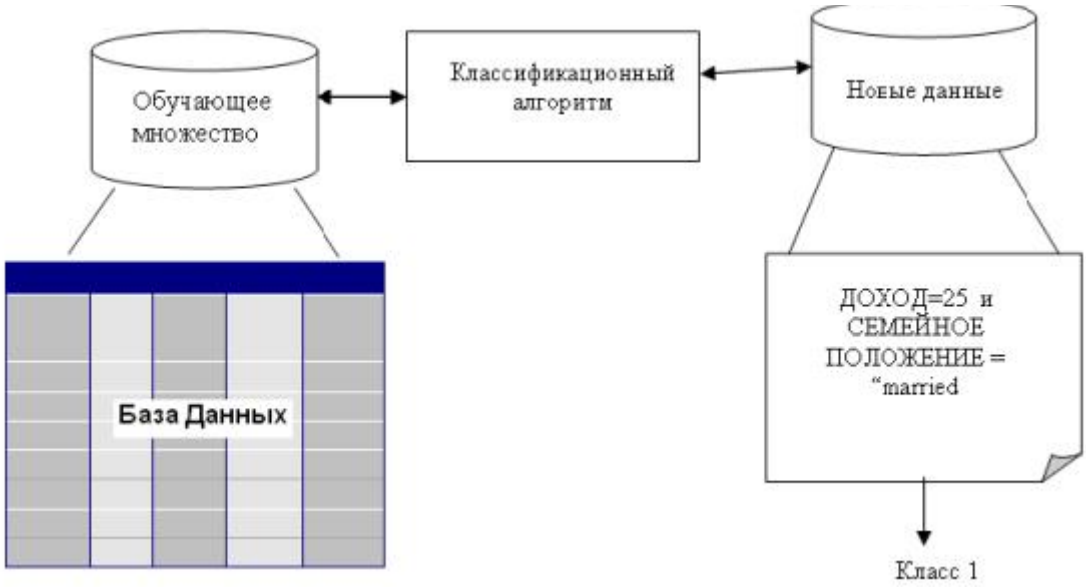

Рис. 5.3. Процесс классификации. Использование модели

Методы, применяемые для решения задач классификации

Для *классификации* используются различные методы. Основные из них:

- *классификация* с помощью деревьев решений;
- байесовская (наивная) *классификация* ;
- *классификация* при помощи *искусственных нейронных сетей*;
- *классификация* методом *опорных векторов*;
- статистические методы, в частности, линейная регрессия;
- *классификация* при помощи метода ближайшего соседа;
- *классификация CBR*-методом;
- *классификация* при помощи *генетических алгоритмов*.

Схематическое решение задачи *классификации* некоторыми методами (при помощи линейной регрессии, деревьев решений и нейронных сетей) приведены на рис. 5.4 - 5.6.

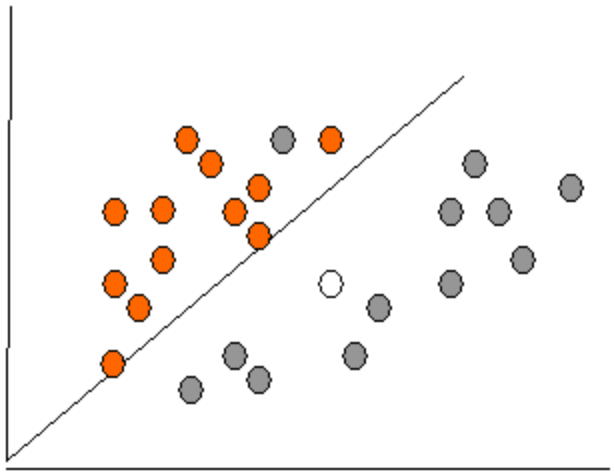

Рис. 5.4. Решение задачи классификации методом линейной регрессии

```
if X > 5 then grey
    else if Y > 3 then orange
            else if X > 2 then grey
                    else orange
```

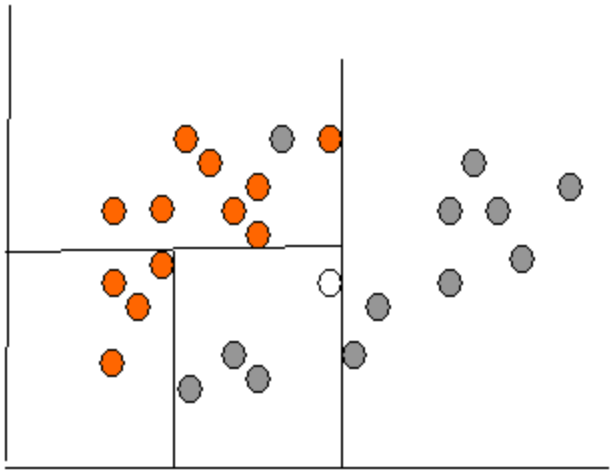

Рис. 5.5. Решение задачи классификации методом деревьев решений

Рис. 5.6. Решение задачи классификации методом нейронных сетей

Точность классификации: оценка уровня ошибок

Оценка точности *классификации* может проводиться при помощи *кросс-проверки*. *Кросс-проверка* (Cross-validation) - это процедура оценки точности *классификации* на данных из тестового *множества*, которое также называют кросс-проверочным множеством. *Точность классификации* тестового *множества* сравнивается с точностью *классификации* обучающего *множества*.

Если *классификация* тестового *множества* дает приблизительно такие же результаты по точности, как и *классификация* обучающего *множества*, считается, что данная модель прошла кросс-проверку.

Разделение на обучающее и тестовое *множества* осуществляется путем деления выборки в определенной пропорции, например обучающее множество - две трети данных и тестовое - одна треть данных. Этот способ следует использовать для выборок с большим количеством примеров. Если же *выборка* имеет малые объемы, рекомендуется применять специальные методы, при использовании которых обучающая и тестовая выборки могут частично пересекаться.

Оценивание классификационных методов

Оценивание методов следует проводить, исходя из следующих характеристик [21]: скорость, *робастность*, интерпретируемость, *надежность*.

Скорость характеризует время, которое требуется на создание модели и ее использование.

Робастность, т.е. *устойчивость* к каким-либо нарушениям исходных предпосылок, означает возможность работы с зашумленными данными и пропущенными значениями в данных.

Интерпретируемость обеспечивает возможность понимания модели аналитиком.

Свойства классификационных правил:

- размер дерева решений;
- компактность классификационных правил.

Надежность методов *классификации* предусматривает возможность работы этих методов при наличии в наборе данных шумов и выбросов.

Задача кластеризации

Только что мы изучили задачу *классификации*, относящуюся к стратегии "*обучение с учителем*".

В этой части лекции мы введем понятия *кластеризации*, *кластера*, кратко рассмотрим классы методов, с помощью которых решается задача *кластеризации*, некоторые моменты процесса *кластеризации*, а также разберем примеры применения кластерного анализа.

Задача *кластеризации* сходна с задачей *классификации*, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены.

Синонимами термина "*кластеризация*" являются "автоматическая *классификация*", "*обучение без учителя*" и "*таксономия*".

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 "сгущений точек".

Цель *кластеризации* - *поиск* существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный *анализ* и изучить "структуру данных".

Само понятие "***кластер***" определено неоднозначно: в каждом исследовании свои "*кластеры*". Переводится понятие *кластер* (*cluster*) как "скопление", "гроздь".

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками *кластера* можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Вопрос, задаваемый аналитиками при решении многих задач, состоит в том, как организовать данные в наглядные структуры, т.е. развернуть таксономии.

Таблица 5.2. Сравнение классификации и кластеризации

Характеристика	**Классификация**	**Кластеризация**
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	*Обучение с учителем*	*Обучение без учителя*
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение	Метки класса обучающего множества неизвестны
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных

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.

В таблице 5.2 приведено сравнение некоторых параметров задач классификации и кластеризации.

На рис. 5.7 схематически представлены задачи классификации и кластеризации.

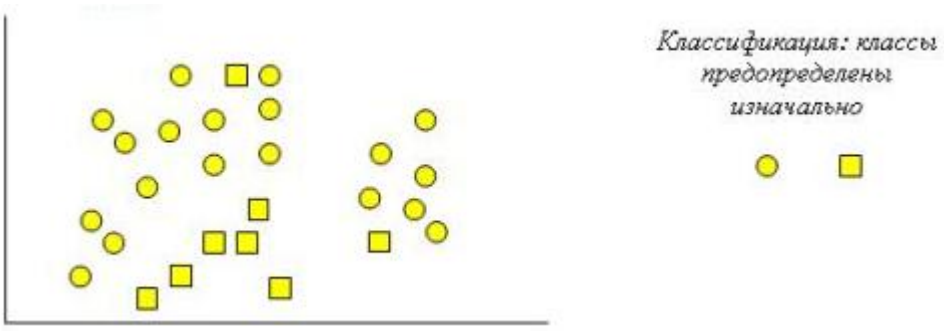

Рис. 5.7. Сравнение задач классификации и кластеризации

Кластеры могут быть непересекающимися, или эксклюзивными (non-*overlapping*, exclusive), и пересекающимися (*overlapping*) [22]. Схематическое изображение непересекающихся и пересекающихся *кластеров* дано на рис. 5.8.

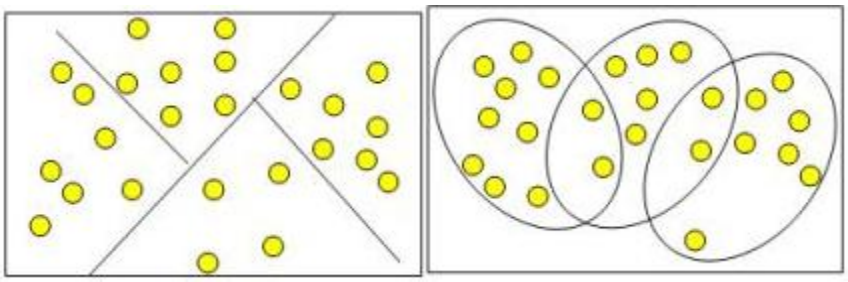

Рис. 5.8. Непересекающиеся и пересекающиеся кластеры

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*.

Подробнее обо всех свойствах кластерного анализа будет рассказано в лекции, посвященной его методам.

На сегодняшний день разработано более сотни различных алгоритмов *кластеризации*. Некоторые, наиболее часто используемые, будут подробно описаны во втором разделе курса лекций.

Приведем краткую характеристику подходов к *кластеризации* [21].

- Алгоритмы, основанные на разделении данных (Partitioning algorithms), в т.ч. итеративные:
  - разделение объектов на k кластеров;
  - итеративное перераспределение объектов для улучшения *кластеризации*.
- Иерархические алгоритмы (*Hierarchy* algorithms):
  - агломерация: каждый объект первоначально является *кластером*, *кластеры*, соединяясь друг с другом, формируют больший *кластер* и т.д.
- Методы, основанные на концентрации объектов (*Density*-based methods):
  - основаны на возможности соединения объектов;
  - игнорируют шумы, нахождение *кластеров* произвольной формы.
- Грид-методы (Grid-based methods):
  - *квантование* объектов в грид-структуры.
- Модельные методы (Model-based):
  - использование модели для нахождения *кластеров*, наиболее соответствующих данным.

Оценка качества кластеризации

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*.

Процесс кластеризации

Процесс *кластеризации* зависит от выбранного метода и почти всегда является итеративным. Он может стать увлекательным процессом и 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*.

Применение кластерного анализа

Кластерный *анализ* применяется в различных областях. Он полезен, когда нужно классифицировать большое *количество информации*. Обзор многих опубликованных исследований, проводимых с помощью кластерного анализа, дал Хартиган (Hartigan, 1975).

Так, в медицине используется *кластеризация* заболеваний, лечения заболеваний или их симптомов, а также *таксономия* пациентов, препаратов и т.д. В археологии устанавливаются таксономии каменных сооружений и древних объектов и т.д. В маркетинге это может быть задача сегментации конкурентов и потребителей. В менеджменте примером задачи *кластеризации* будет *разбиение* персонала на различные группы, *классификация* потребителей и поставщиков, выявление схожих производственных ситуаций, при которых возникает брак. В медицине - *классификация* симптомов. В социологии задача *кластеризации* - *разбиение* респондентов на однородные группы.

Кластерный анализ в маркетинговых исследованиях

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.

Так, 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. Эта проблема подробно описана в работах Клакстона, Фрая и Портиса (1974), Киля и Лэйтона (1981).

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 компании. 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.

Практика применения кластерного анализа в маркетинговых исследованиях

Приведем некоторые известные статьи, посвященные применению кластерного анализа для маркетинговых исследований.

В 1971 году была опубликована статья о сегментации клиентов по сфере интересов на основе данных, характеризующих предпочтения клиентов.

В 1974 году была опубликована статья Секстона (Sexton), целью которой была идентификация групп семей - потребителей продукта, в результате были разработаны стратегии позиционирования бренда. Основой для исследований были рейтинги, которые респонденты присваивали продуктам и брендам.

В 1981 году была опубликована статья, где проводился анализ поведения покупателей новых автомобилей на основе данных факторных нагрузок, полученных при анализе набора переменных.

Выводы

В этой лекции нами были подробно рассмотрены задачи *классификации* и *кластеризации*. Несмотря на кажущуюся похожесть этих задач, решаются они разными способами и при помощи разных методов. Различие задач прежде всего в исходных данных.

Классификация, являясь наиболее простой задачей *Data* Mining, относится к стратегии "*обучение с учителем*", для ее решения обучающая *выборка* должна содержать значения как входных переменных, так и выходных (целевых) переменных. *Кластеризация*, напротив, является задачей *Data* Mining, относящейся к стратегии "*обучение без учителя*", т.е. не требует наличия значения целевых переменных в обучающей выборке.

Задача *классификации* решается при помощи различных методов, наиболее простой - линейная регрессия. Выбор метода должен базироваться на исследовании исходного набора данных. Наиболее распространенные методы решения задачи *кластеризации*: метод k-средних (работает только с числовыми атрибутами), иерархический кластерный *анализ* (работает также с символьными атрибутами), метод *SOM*. Сложностью *кластеризации* является необходимость ее оценки.