**Основные задачи Data Mining. Какие дисциплины охватывает Data Mining?**

Накопление данных ->анализ данных -> извлечение знаний= когнитивный процесс

Data Mining — это процесс автоматического извлечения интересных и ранее неизвестных знаний из больших объемов данных. Основная задача Data Mining заключается в поиске скрытых закономерностей, тенденций, корреляций, классификации и кластеризации данных.

Дисциплины, охватываемые Data Mining, включают в себя:

1.Статистика - для анализа и интерпретации данных

2.Искусственный интеллект - для создания алгоритмов машинного обучения и интеллектуального анализа данных

3.Базы данных - для хранения и организации больших объемов данных

4.Математика - для разработки математических моделей и алгоритмов для анализа данных

5.Визуализация данных - для представления и визуального анализа данных.

Data Mining также имеет практическое применение в различных областях, включая маркетинг, банковское дело, медицину, финансы, производство и другие.

Задачи Data Analysis

1 Классификация

2 Кластеризация

3 Прогнозирование

4 Ассоциация

5 Обнаружение и анализ и отклонений

6 Оценивание

7 Анализ связей

8 Визуализация

9 Подведение итогов (Summarization)

**Классификация методов DM.**

Часто задачи Data Analysis подразделяют на следующие:

•исследования и открытия,

•прогнозирования и классификации,

•объяснения и описания.

•Метод (method) представляет собой норму или правило,определенный путь, способ, прием решений задачи теоретического, практического, познавательного,управленческого характера.

•Алгоритм (algorithm) - точное предписание относительно последовательности действий (шагов), преобразующих исходные данные в искомый результат.

Классификация методов Data Analysis:

1. Технологические

2. Статистические

3. Кибернетические

Технологические методы

• Непосредственное использование данных, или сохранение данных:

– кластерный анализ, метод k-ближайшего соседа, рассуждение по аналогии.

• Выявление и использование формализованных закономерностей (дистилляция шаблонов):

– методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях.

– логические методы, или методы логической индукции, включают нечеткие запросы и анализы, символьные правила, деревья решений, генетические алгоритмы.

Методы этой группы хорошо подходят для интерпретации.

Статистические методы

• Дескриптивный анализ и описание исходных данны.

• Анализ связей (корреляционный и регрессионный анализ,факторный анализ, дисперсионный анализ).

• Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).

• Анализ временных рядов (динамические модели и прогнозирование).

Методы этой группы тоже хорошо подходят для интерпретации.

Кибернетические методы

Искусственные нейронные сети (распознавание,кластеризация, прогноз);

Эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);

Генетические алгоритмы (ГА-оптимизация);

Ассоциативная память (поиск аналогов, прототипов);

Нечеткая логика;

Деревья решений;

Системы обработки экспертных знаний.

Методы этой группы являются эмпирическими и хорошо подходят для прогнозирования и распознавания.

Свойства методов Data Analysis

• Различные методы Data Analysis характеризуются определенными свойствами, которые могут быть определяющими при выборе метода анализа данных. Методы можно сравнивать между собой, оценивая характеристики их свойств.

• Среди основных свойств и характеристик методов Data Analysis выделим следующие:

точность, масштабируемость, интерпретируемость,проверяемость, трудоемкость, гибкость, скорость выполнения и популярность.

**Понятие кластерного анализа, Классификация алгоритмов кластеризации.**

Кластеризация (или кластерный анализ) — задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных групп должны быть как можно более отличны. Перечень групп четко не задан и определяется в процессе работы алгоритма.

Виды алгоритмов кластеризации:

I. Иерархические и плоские – по способу организации

Иерархические алгоритмы строят систему вложенных разбиений. Т.е. на выходе мы получаем дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластера.

Плоские алгоритмы строят одно разбиение объектов на кластеры, т.е формируют кластеры одного ранга.

II. Четкие и нечеткие – по способу определения принадлежности

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. Т.е. каждый объект относится к каждому кластеру с некоторой вероятностью.

1. Алгоритмы, разбивающие данные на заданное число кластеров (то есть число кластеров – это входной параметр алгоритма). Пример: алгоритм k-means.

2. Алгоритмы, в которых число кластеров не определено заранее, а вычисляется самим алгоритмом. Пример: алгоритм FOREL.

На основе дополнительных характеристик можно выделить:

• алгоритмы, оперирующие мерами близости/расстояния объектов множества;

• алгоритмы, работающие с векторами характеристик;

• вероятностные алгоритмы, основанные на порождающей модели (например, НММ-кластеризация семейств последовательностей);

• граф-ориентированные — предназначенные для анализа информации, представленной в виде сетей

• бикластеризации — обработка данных серии экспериментов над множеством образцов, заключающаяся в выделении наборов строк и столбцов матрицы результатов, с определенными паттернами распределения значений

**Зачем нужна мера близости в кластеризации? В чем достоинства алгоритмов, построенных на основе теории графов? Перечислите виды алгоритмов кластеризации на графах.**

Мера близости в кластеризации используется для измерения сходства между объектами в выборке и группировки их в кластеры на основе этой меры. Она позволяет определить, насколько похожи два объекта между собой, и разбить все объекты на кластеры, где объекты внутри кластера имеют высокую степень сходства, а объекты между кластерами имеют низкую степень сходства.

Алгоритмы, основанные на теории графов (иерархические)

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа G=(V, E), вершинам которого соответствуют объекты, а ребра имеют. вес, равный «расстоянию между объектами. Достоинством графовых алгоритмов кластеризации являются наглядность, относительная простота реализации и возможность внесения различных усовершенствований, основанных на геометрических соображениях. Основными алгоритмами являются алгоритм выделения связных компонент, алгоритм построения минимального покрывающего (остовного) дерева и алгоритм послойной кластеризации.

**В чем суть алгоритмов нахождения квадратичной ошибки?**

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения: e2 (X,L)=j=0Ki=0njx(j)i -cj2 . где с — «центр масс» кластера j (точка со средними значениями характеристик для данного кластера), K – количество кластеров, n – количество точек.

Алгоритмы квадратичной ошибки относятся к типу плоских алгоритмов. Самым распространенным алгоритмом этой категории является метод k-средних. Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга.

Работа алгоритма делится на несколько этапов:

1. Случайно выбрать k точек, являющихся начальными «центрами масс» кластеров.

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

3. Пересчитать «центры масс» кластеров согласно их текущему составу.

4. Если критерий остановки алгоритма не удовлетворен, вернуться к п.

2. В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Также возможно останавливать работу алгоритма, если на шаге 2 не было объектов, переместившихся из кластера в кластер.

К недостаткам данного алгоритма можно отнести необходимость задавать количество кластеров для разбиения.

**Плоские алгоритмы кластеризации перечислить, охарактеризовать работу.**

Виды алгоритмов кластеризации:

Иерархические и плоские – по способу организации.

Плоские алгоритмы строят одно разбиение объектов на кластеры, т.е формируют кластеры одного ранга.

Алгоритмы плоской кластеризации:

1 Алгоритмы квадратичной ошибки

1.1 Метод k-средних

1.2 Метод k -ближайших соседей

2 Нечеткие алгоритмы

2.1 Метод с-средних

1.1 - Метод k-средних

Этот алгоритм строит заданное число кластеров, расположенных как можно дальше друг от друга. Работа алгоритма делится на несколько этапов:

1.Случайно выбрать k точек, являющихся начальными «центрами масс» кластеров.

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

3 Пересчитать «центры масс» кластеров согласно их текущему составу.

Если критерий остановки алгоритма не удовлетворен, вернуться к п. 2

В качестве критерия остановки работы алгоритма обычно выбирают минимальное изменение среднеквадратической ошибки. Также возможно останавливать работу

алгоритма, если на шаге 2 не было объектов,переместившихся из кластера в кластер.

1.2 - Метод k -ближайших соседей (k-nearest neighbors algorithm, k-NN) —метрический алгоритм для автоматической классификации объектов или регрессии.

В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны.

Алгоритм может быть применим к выборкам с большим количеством атрибутов (многомерным). Для этого перед применением нужно определить функцию расстояния; классический вариант такой функции — евклидова метрика.

2.1 Метод с-средних - Алгоритм является модификацией метода k-средних.

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

В целом основная идея плоских алгоритмов кластеризации заключается в том, чтобы разбить данные на группы, так чтобы объекты внутри одной группы были как можно более похожими друг на друга, а объекты из разных групп были как можно более различными.

**Поясните суть работы алгоритмов нахождения связных компонент и алгоритмов покрывающего дерева.**

Нахождение связанных компонент:

Изначально задается входной параметр R и из графа удаляются все ребра, для которых “расстояния” больше чем R.

Суть алгоритма - подобрать такое R, лежащее в диапазоне всех расстояний, при котором граф разбивается на несколько связных компонент. Полученные компоненты и есть кластеры.

Для подбора параметра R обычно строится гистограмма распределений попарных расстояний. В задачах с хорошо выраженной кластерной структурой данных на гистограмме будет два пика – один соответствует внутрикластерным расстояниям, второй – межкластерным расстояния.

Параметр R подбирается из зоны минимума между этими пиками. При этом управлять количеством кластеров при помощи порога расстояния довольно затруднительно.

Минимальное покрывающее дерево:

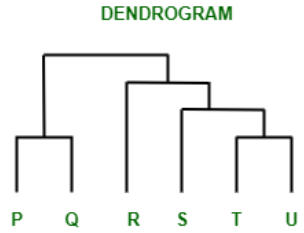
Строим на графе минимальное покрывающее дерево, а затем последовательно удаляем ребра с наибольшим весом. При удалении таких ребер граф разбивается на части, которые являются компонентами

**Опишите шаги построения дендрограммы.**

Дендрограмма - это древовидная диаграмма, используемая для визуализации взаимосвязи между кластерами. Чем больше расстояние между вертикальными линиями на дендрограмме, тем больше расстояние между этими кластерами. Ключом к интерпретации дендрограммы является концентрация на высоте, на которой любые два объекта соединены вместе.

Пример дендрограммы:

Предположим, у нас есть шесть кластеров: P, Q, R, S, T и U. Иерархическая дендрограмма кластеров из этих шести наблюдений, показанная на диаграмме рассеяния, имеет вид:



Как выполнить?

Каждая точка данных должна рассматриваться как кластер в начале. Обозначим количество кластеров в начале как K.

Сформируйте один кластер, объединив две ближайшие точки данных, в результате чего получится K-1 кластер. Значение нового кластера - среднее между значениями его родителей.

Повторяйте 3-й шаг, пока не будет создан один большой кластер.

**В чем состоит суть стандартизации и нормализации переменных, зачем они нужны?**

Стандартизация и нормализация переменных являются методами предварительной обработки данных и используются для приведения значений переменных к более удобному и понятному формату.

Суть стандартизации заключается в том, что значения переменных масштабируются таким образом, чтобы они имели среднее значение равное 0 и стандартное отклонение равное 1. Это позволяет сравнивать значения переменных с разными единицами измерения на более равных условиях и уменьшает влияние выбросов на результаты анализа.

Нормализация заключается в приведении значений переменных к диапазону от 0 до 1. Так можно сравнивать значения переменных, которые сильно отличаются друг от друга.