ЛЕКЦИЯ 7
МЕТОДЫ КЛАСТЕРИЗАЦИИ И СНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ

Цель лекции

Изучить основные методы обучения без учителя, применяемые для выявления скрытой структуры данных и компактного представления многомерных объектов; рассмотреть алгоритмы K-Means, иерархическую кластеризацию и DBSCAN, а также методы PCA и t-SNE и особенности их применения в задачах искусственного интеллекта и информационной безопасности.

Обучение без учителя

В задачах обучения без учителя данные не содержат заранее заданной целевой переменной с правильными ответами. Алгоритм анализирует признаки объектов и пытается выявить внутреннюю структуру данных. К основным задачам относятся кластеризация, снижение размерности и поиск аномалий.

Задача	**Назначение**	**Пример**
Кластеризация	Объединение похожих объектов в группы	Группировка сетевых соединений по характеру поведения
Снижение размерности	Получение компактного представления большого числа признаков	Сжатие многомерных признаков трафика до нескольких компонент
Выявление аномалий	Поиск объектов, отличающихся от основной структуры данных	Обнаружение необычной активности пользователя

Понятие кластеризации

Кластеризация — метод анализа данных, при котором объекты объединяются в группы, называемые кластерами. Желательно, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты разных кластеров — различались.

В отличие от классификации, номера или названия кластеров заранее неизвестны. Алгоритм самостоятельно формирует группы на основании структуры признакового пространства.

Пример. Имеется набор событий информационной системы, описанных частотой обращений, объёмом переданных данных, временем активности и количеством неуспешных попыток входа. Кластеризация может выявить группы типичного поведения и небольшую группу необычных событий, требующую дополнительного анализа.

Метрики расстояния и масштабирование

Многие методы кластеризации используют понятие расстояния между объектами. Одной из наиболее распространённых является евклидова метрика:

$$d(x, y) = \sqrt{\Sigma(x_i - y_i)^2}$$

Евклидово расстояние зависит от масштаба признаков. Если один признак изменяется от 0 до 1, а другой — от 0 до 100000, второй признак будет существенно сильнее влиять на расстояние. Поэтому перед K-Means и рядом других методов данные обычно стандартизируют или нормализуют.

Метрика	Идея	Особенность
Евклидова	Прямое геометрическое расстояние	Широко применяется в K-Means
Манхэттенская	Сумма абсолютных разностей	Может быть устойчивее к отдельным большим отклонениям
Косинусная мера	Сравнение направления векторов	Часто применяется к текстовым векторам

Алгоритм K-Means

K-Means — один из наиболее известных алгоритмов кластеризации. Пользователь заранее задаёт число кластеров K. Алгоритм ищет K центров — центроидов — и распределяет объекты между ними.

Основные шаги алгоритма:

1. Выбрать число кластеров K.
2. Инициализировать K центроидов.
3. Для каждого объекта вычислить расстояние до центроидов и назначить объект ближайшему кластеру.
4. Пересчитать каждый центроид как среднее положение объектов его кластера.
5. Повторять назначение объектов и пересчёт центроидов до стабилизации решения.

Цель K-Means — уменьшить суммарное внутрикластерное расстояние объектов до соответствующих центроидов. В литературе эта величина часто называется inertia или within-cluster sum of squares.

Пример. Для анализа активности пользователей можно использовать признаки: число входов, средняя длительность сессии, объём переданных данных и количество ошибок аутентификации. K-Means может разделить пользователей на несколько поведенческих групп.

Выбор числа кластеров K

Одной из главных особенностей K-Means является необходимость заранее определить K. На практике используют сочетание количественных критериев и содержательной интерпретации.

Подход	Смысл
Метод локтя	Анализируется уменьшение внутрикластерной ошибки при росте K; выбирается область, после которой улучшение становится менее существенным
Silhouette score	Сравнивается близость объекта к своему кластеру и удалённость от соседних кластеров
Экспертная интерпретация	Проверяется, имеют ли найденные группы практический смысл

Нельзя выбирать K только по красивой визуализации. Результат должен быть устойчивым и интерпретируемым в контексте конкретной задачи.

Преимущества и ограничения K-Means

Преимущества	Ограничения
Простота и высокая скорость	Необходимо заранее задавать K
Хорошо работает с компактными кластерами	Чувствителен к масштабу признаков

Подходит для больших наборов данных	Чувствителен к выбросам
Центроиды удобно интерпретировать	Сложные кластеры произвольной формы описывает хуже

Иерархическая кластеризация

Иерархическая кластеризация строит вложенную структуру групп. При агломеративном подходе каждый объект сначала рассматривается как отдельный кластер, после чего наиболее близкие кластеры последовательно объединяются. Результат можно представить в виде дендрограммы.

Иерархический подход полезен, когда необходимо увидеть структуру данных на нескольких уровнях детализации и исследовать, какие объекты объединяются раньше, а какие — только на более высоком уровне.

При объединении кластеров необходимо определить linkage — способ вычисления расстояния между группами. Используются single linkage, complete linkage, average linkage и метод Ward.

Алгоритм DBSCAN

DBSCAN — плотностный метод кластеризации. В отличие от K-Means он не требует заранее задавать число кластеров и способен выделять области произвольной формы. Алгоритм рассматривает плотность объектов в окрестности каждой точки.

Ключевые параметры DBSCAN — eps, определяющий радиус окрестности, и min_samples, задающий минимальное количество объектов, необходимое для формирования плотной области.

Тип точки	**Смысл**
Core point	В её окрестности находится достаточно объектов
Border point	Находится рядом с плотной областью, но сама не удовлетворяет условию ядра
Noise	Не относится к обнаруженным плотным областям

Способность помечать отдельные объекты как шум делает DBSCAN интересным для первичного поиска необычных событий. Однако обнаруженный шум нельзя автоматически считать атакой: это только объект, не вписавшийся в плотностную структуру данных.

Сравнение методов кластеризации

Метод	**Нужно задавать число кластеров**	**Форма кластеров**	**Работа с шумом**	**Основная особенность**
K-Means	Да	Предпочтительно компактная	Слабая	Быстрый и простой
Иерархическая	Не обязательно на первом этапе	Различная	Зависит от настройки	Показывает иерархию групп
DBSCAN	Нет	Произвольная	Хорошая	Выделяет плотные области и шум

Оценка качества кластеризации

В кластеризации правильные ответы часто отсутствуют, поэтому оценка отличается от классификации. Используются внутренние критерии, основанные на компактности и разделимости групп, а также экспертная интерпретация.

Критерий	**Интерпретация**
Inertia	Насколько объекты близки к центрам своих кластеров; используется при анализе K-Means
Silhouette score	Сочетает компактность своего кластера и отделённость от соседних
Стабильность	Насколько структура сохраняется при изменении выборки или инициализации
Интерпретируемость	Можно ли содержательно объяснить различия между найденными группами

Кластеризация в информационной безопасности

В информационной безопасности кластеризация может применяться для анализа поведения пользователей, группировки сетевых потоков, исследования событий SIEM, сегментации вредоносных образцов и поиска необычных паттернов.

Например, сетевые потоки можно представить признаками duration, packet_count, bytes_in, bytes_out, average_packet_size и connection_rate. После стандартизации K-Means формирует группы похожих потоков. Затем аналитик изучает характеристики кластеров и определяет, соответствуют ли они типичному поведению или требуют дополнительной проверки.

В материалах по информационной безопасности также используются K-Means и DBSCAN как методы аномалийно-ориентированного анализа сетевых и хостовых метрик. Их преимущество состоит в возможности работать без заранее подготовленных меток, а ограничением является риск ложных срабатываний и необходимость корректно описывать нормальное поведение.

Что такое снижение размерности

Современные наборы данных могут содержать десятки, сотни и тысячи признаков. Большая размерность увеличивает вычислительные затраты, затрудняет визуализацию и может приводить к избыточности признаков. Снижение размерности преобразует исходное пространство в пространство меньшей размерности с сохранением наиболее существенной информации.

Основные цели снижения размерности: визуализация многомерных данных; уменьшение вычислительной сложности; устранение избыточности и коррелированных признаков; подготовка компактного представления для последующего анализа.

Метод главных компонент PCA

Principal Component Analysis — PCA — линейный метод снижения размерности. Он преобразует исходные признаки в новые взаимно ортогональные направления — главные компоненты. Первая компонента описывает максимально возможную долю вариации данных, следующая — максимальную оставшуюся вариацию при условии ортогональности предыдущей и т. д.

PCA не просто удаляет исходные столбцы. Он создаёт новые признаки как линейные комбинации исходных переменных.

Общая последовательность применения PCA включает стандартизацию признаков, построение ковариационной структуры, определение главных направлений и проекцию объектов на выбранное число компонент.

Объяснённая дисперсия

При использовании PCA необходимо определить, сколько компонент оставить. Для этого анализируется explained variance ratio — доля дисперсии, объясняемая каждой компонентой. Суммарная объяснённая дисперсия показывает, какая часть вариативности исходных данных сохраняется после преобразования.

Например, если первые 10 компонент сохраняют большую часть объяснённой дисперсии исходных 100 признаков, их можно использовать как более компактное представление. Конкретный порог сохранения информации выбирается в зависимости от задачи, а не как универсальное значение.

PCA для текстовых и высокоразмерных признаков

В загруженных материалах PCA применяется к числовому представлению текстов TF-IDF. Каждая строка такой матрицы соответствует документу, а столбцы — терминам. Число признаков может быть очень большим, поэтому снижение размерности позволяет сформировать компактные латентные компоненты.

После преобразования анализируется explained_variance_ratio_, чтобы оценить объём информации, сохранённый в сокращённом пространстве. Это позволяет выбирать число компонент как компромисс между компактностью и сохранением структуры данных.

Метод t-SNE

t-Distributed Stochastic Neighbor Embedding — t-SNE — нелинейный метод снижения размерности, часто применяемый для визуализации многомерных данных в двух- или трёхмерном пространстве.

Главная задача t-SNE — сохранить локальное соседство: объекты, близкие в исходном многомерном пространстве, стремятся располагаться близко и на визуализации. Поэтому t-SNE удобен для исследовательского анализа возможных групп.

В отличие от PCA, координаты t-SNE обычно не интерпретируют как отдельные содержательные признаки. Расстояния между далёкими группами и размеры визуальных кластеров также не следует трактовать слишком буквально.

PCA и t-SNE: различия

Характеристика	**PCA**	**t-SNE**
Тип метода	Линейный	Нелинейный
Основная цель	Компактное представление и сохранение дисперсии	Визуализация локальной структуры
Новые признаки	Главные компоненты	Координаты визуального вложения
Интерпретация	Через компоненты и explained variance	В основном визуальная
Применение к новым объектам	Удобно после обучения преобразования	Не является основной сильной стороной классического t-SNE

Связь кластеризации и снижения размерности

Эти методы часто используются совместно. Сначала данные могут быть стандартизированы и преобразованы PCA, затем в компактном пространстве выполняется кластеризация. Другой вариант — сначала получить кластеры, а затем использовать PCA или t-SNE для двумерной визуализации результата.

Важно помнить, что визуальная картинка после t-SNE сама по себе не доказывает существование реальных кластеров. Кластеризация и визуализация должны оцениваться отдельными критериями.

Пример K-Means в Python

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

X = data[['duration', 'packets', 'bytes', 'failed_logins']]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

model = KMeans(n_clusters=3, random_state=42, n_init='auto')
labels = model.fit_predict(X_scaled)

print('Silhouette:', silhouette_score(X_scaled, labels))
print('Centers:', model.cluster_centers_)
```

Пример DBSCAN в Python

```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import DBSCAN

X_scaled = StandardScaler().fit_transform(X)

dbscan = DBSCAN(eps=0.6, min_samples=5)
labels = dbscan.fit_predict(X_scaled)

data['cluster'] = labels
print(data['cluster'].value_counts())
```

21. Пример PCA в Python

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

X_scaled = StandardScaler().fit_transform(X)

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

print('Explained variance:', pca.explained_variance_ratio_)
print('Total:', pca.explained_variance_ratio_.sum())
```

Типичные ошибки

– Применять K-Means к немасштабированным признакам с сильно различающимися диапазонами.

– Выбирать число кластеров только по визуальному впечатлению.

– Считать номера кластеров готовыми смысловыми классами.

– Автоматически интерпретировать все точки DBSCAN с меткой noise как атаки.

– Использовать t-SNE как доказательство наличия кластеров.

– Выбирать число компонент PCA без анализа объяснённой дисперсии.

– Оценивать кластеризацию только одним числовым критерием без анализа содержимого групп.

Практическая схема выбора метода

Ситуация	Подход
Нужно быстро разделить объекты на заданное число компактных групп	K-Means
Число групп заранее неизвестно и важен поиск плотных областей и шума	DBSCAN
Нужно исследовать иерархическую структуру объектов	Иерархическая кластеризация
Нужно компактное линейное представление признаков	PCA
Нужно визуально исследовать локальную структуру многомерных данных	t-SNE

Основные выводы

– Кластеризация относится к обучению без учителя и выявляет группы без заранее заданных меток.

– K-Means формирует K кластеров вокруг центроидов и требует выбора числа кластеров.

– DBSCAN выделяет плотные области и шум и не требует заранее задавать число кластеров.

– Результат кластеризации зависит от признаков, масштаба данных, метрики расстояния и параметров алгоритма.

– PCA формирует новые главные компоненты и позволяет уменьшить размерность с контролем объяснённой дисперсии.

– t-SNE прежде всего применяется для визуального исследования локальной структуры многомерных данных.

– В задачах информационной безопасности методы без учителя полезны для исследования поведения и поиска необычных паттернов, но требуют экспертной интерпретации.

Контрольные вопросы

1. Что понимается под обучением без учителя?
2. Чем кластеризация отличается от классификации?
3. Что такое кластер?
4. Какую роль играет метрика расстояния?
5. Почему перед K-Means важно масштабировать признаки?
6. Каковы основные шаги K-Means?
7. Что такое центроид?
8. Почему в K-Means необходимо задавать K?
9. В чём состоит метод локтя?
10. Что показывает silhouette score?

11. Какие ограничения имеет K-Means?
12. Как работает иерархическая кластеризация?
13. Что показывает дендрограмма?
14. Что такое linkage?
15. В чём состоит идея DBSCAN?
16. Что означают параметры eps и min_samples?
17. Какие точки DBSCAN считает шумом?
18. Почему шум DBSCAN нельзя автоматически считать атакой?
19. Как оценивать качество кластеризации без известных меток?
20. Как кластеризация применяется в информационной безопасности?
21. Что понимается под снижением размерности?
22. Какие задачи решает PCA?
23. Что такое главная компонента?
24. Что показывает explained variance ratio?
25. Почему PCA полезен для высокоразмерных данных?
26. Для чего используется t-SNE?
27. Чем PCA отличается от t-SNE?
28. Почему визуализация t-SNE не доказывает наличие кластеров?
29. Как совместно применять PCA и кластеризацию?
30. Как выбрать метод для конкретной задачи?

Список источников:

1. Силлабус дисциплины «Искусственный интеллект» для образовательной программы 7M06301 «Комплексное обеспечение информационной безопасности». Тема недели 7 — «Методы кластеризации и снижения размерности данных».

2. Учебное пособие «Основы искусственного интеллекта». Глава 7 «Кластеризация и снижение размерности»: K-Means, DBSCAN, PCA, t-SNE.

3. Материалы по применению TF-IDF, PCA и t-SNE для снижения размерности текстовых признаков и анализа объяснённой дисперсии.

4. Материалы проекта по анализу судебных данных: K-Means, иерархическая кластеризация и DBSCAN для выявления скрытых структур и групп без предварительной разметки.