ДИСЦИПЛИНА	**Проектирование интеллектуальных систем (часть 1/2)**
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**вычислительной техники**
ВИД УЧЕБНОГО МАТЕРИАЛА	**Материалы для практических/семинарских занятий**
ПРЕПОДАВАТЕЛЬ	**Холмогоров Владислав Владиславович**
СЕМЕСТР	**6, 2023-2024**

Практическая работа № 2

«Кластерный анализ данных»

по дисциплине «Проектирование интеллектуальных систем (часть 1/2)»

Цели: приобрести навыки кластеризации как инструмента исследовательского/разведочного анализа данных (exploratory data analysis).

Задачи:

1) Провести кластеризацию исходного набора данных, выполнив следующие шаги:

– определить предметную область решаемой задачи, ею могут выступать сферы продаж, производства, анализ социальных систем, интеллектуальное планирование, анализ природных, биологических и медицинских данных, анализ трафика и общая кибербезопансоть, финансовый и маркетинговый анализ, анализ в сфере производства и т.д.;

– найти или сгенерировать набор данных для выбранной задачи, проведя предварительную предобработку и подготовку данных (см. Примечание 1);

– визуализировать подготовленные данные и провести предиктивную аналитику количества и качества кластеров (см. Примечание 2);

– выполнить кластеризацию подготовленного набора данных хотя бы одним из методов любой категории алгоритмов (см. Примечание 3).

2) В качестве дополнительного задания реализовать хотя бы один из следующих пунктов (один на оценку 4, два и более – на оценку 5):

– выполнить кластеризацию не менее чем двумя методами разных категорий (см. Примечание 3), сравнить и описать результаты;

– интерпретировать полученные кластеры (см. Примечание 4);

– определить качество кластеризации на основе соответствующих метрик (см. Примечание 5)

– выполнить с помощью кластеризации очистку данных от шумов как один из этапов предобработки данных (см. Примечание 6).

ПРИМЕЧАНИЕ:

1) При проведении кластеризации необходимо определить её цели, критерии и исследуемые характеристики; провести минимальную предобработку данных, в которую входят очистка от пропусков и ошибок, а также строк-дубликатов; хотя кластерный анализ сам по себе может быть применён при подавлении шумовых данных, имеет смысл заранее убрать аномальные значения, также имеет смысл отобрать признаки (как правило, с помощью корреляционного анализа и методов понижения размерности).

2) Так как некоторые методы кластеризации требуют указания количества кластеров, визуализация набора данных с целью самостоятельной оценки возможного их количества и качества является необходимой мерой, но даже для методов, не требующих указания количества кластеров (вроде иерархической кластеризации, DBSCAN-а и модели гауссовой смеси) визуализация является важным компонентом решения, т.к. позволяет самостоятельно оценить результаты работы алгоритма.

3) Методы кластеризации можно разделить на следующие категории:

- на основе группы/разделения/центроидов: k-means, K-medoids, PAM, CLARA, CLARANS, CluStream, Fuzzy C-means и Hierarchical K-means;
- методы на основе плотности: DBSCAN, DBCLASD, OPTICS, denpro, kNN Density Estimation, DENCLUE, DenStream;
- на основе сетки: STING, WaveCluster, CLIQUE, OptiGrid, D-stream, Axis Shifted Grid Clustering Algorithm (ASGC);
- иерархические методы: ROCK, CURE, BIRCH, Echidne, Chamelion, ClusTree, агломеративная кластеризация, разделительная иерархическая кластеризация;
- методы на основе ограничений: COP-K-Means, CONstraint-baseb-CLUSteringn, PCKmeans, constraint spectral partitioning clustering, CMWK-Means;

– модельно-ориентированные: модель гауссовской смеси, EM, COBWEB, CLASSIT, SOMs, SWEM, Latent class model, статистический подход, extension model based k-means;

4) Кластеры – ещё неизвестные классы, поэтому визуальный и смысловой анализы могут позволить интерпретировать группы, на которые алгоритм кластеризации на основе признаков разбил набор данных, если для работы с кластеризацией и классификацией использовать один и тот же набор данных, из которого для кластеризации удалить метки классов, то можно напрямую сравнить результаты кластеризации с классификацией.

5) Метрики кластеризации обычно разделяются на группы – внешние (Extrinsic Measures) и внутренние (Intrinsic Measures), первые содержат метрики: скорректированный индекс Рэнда, случайный индекс, энтропия, оценки Фаулкса-Мэллоуза, оценки на основе взаимной информации, Homogeneity, Completeness и V-меру; вторые содержат метрики: Compactness, Separation, индекс Дэвиса-Болдина, индекс валидности Думма, сумма квадратичной ошибки, Silhouette.

6) Некоторые методы кластеризации (особенно методы, основанные на плотности точек) воспринимают отдельно стоящие точки как шумовые и не учитывают их при формировании кластеров, что позволяет автоматически избавится от шумовых и аномальным значений в наборе данных, такой подход можно использовать для очистки данных для другой модели, которая будет обучаться и работать на этих данных, либо это свойство шумоустойчивости просто будет фактически присутствовать при использовании методов.

ОСОБЫЙ БОНУС (доступен только в том случае, если выполнены пункты 2-го задания):

Если в первой практической работе был выполнен кластерный анализ с помощью методов анализа ассоциативных правил, то провести на том же наборе данных (в таком случае исходный набор данных должен подходить по

признакам для обеих задач, если его не получается найти, то его можно сгенерировать или получить методами интеграции данных) кластеризацию традиционными методами и сравнить результаты.