Министерство науки и высшего образования РФ

Федеральное государственное бюджетное

образовательное учреждение высшего образования

ВЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет автоматики и вычислительной техники

Кафедра радиоэлектронных средств

Отчет по лабораторному практикуму №5

**Анализ научных методов**

дисциплина «Методология научных исследований»

|  |  |
| --- | --- |
| Выполнил: студент группы ИТм-1301 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Иванов И.И. / |
|  |  |
| Проверил: профессор кафедры РЭС | \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Частиков А.В. / |

Киров 2022

**Цель**: изучить сущность научных методов, выявить их достоинства и недостатки.

**Задачи**:

1. Выявить сущность для заданного преподавателем научного метода.

2. Кратко сформулировать сущность научного метода.

4. Выявить применимость, достоинства и недостатки, действенность данного научного метода для изучаемой области знаний (темы научных исследований).

5. Определить антипод научного метода.

6. Выполнить пп. 2-4 для антипода.

7. Составить отчет в электронной форме. В отчете сделать обобщенные развернутые выводы.

**1 Формулировка сущности научного метода**

Тема научных исследований: …………………………………………………………………….

Для данной темы научных исследований лучше всего подходит научный метод – “Кластерный анализ” потому что, (аргументировать):

а) …………………………………………………………………………………………………..;

б) …………………………………………………………………………………………………..;

………………………………………………………………………………………………………

Кластерный анализ (англ. *cluster analysis*) – многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы [ ]. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Спектр применений кластерного анализа очень широк: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, социологии, геологии и других дисциплинах. Однако универсальность применения привела к появлению большого количества несовместимых терминов, методов и подходов, затрудняющих однозначное использование и непротиворечивую интерпретацию кластерного анализа.

Кластерный анализ выполняет следующие основные задачи:

* разработка типологии или классификации;
* исследование полезных концептуальных схем группирования объектов;
* порождение гипотез на основе исследования данных;
* проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

* отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные;
* определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства;
* вычисление значений той или иной меры сходства (или различия) между объектами.
* применение метода кластерного анализа для создания групп сходных объектов;
* проверка достоверности результатов кластерного решения.

Два фундаментальных требования, предъявляемых к данным – однородность и полнота. Однородность требует, чтобы все кластеризуемые сущности были одной природы, описывались сходным набором характеристик.

Цели кластеризации:

– понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»);

– сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера;

– обнаружение новизны (англ. novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд групп подходов (некоторые методы можно отнести сразу к нескольким группам и потому необходимо рассматривать данную типизацию как некоторое приближение к реальной классификации методов кластеризации).

Методы кластеризации:

1. Вероятностный подход. Предполагается, что каждый рассматриваемый объект относится к одному из k классов. Некоторые авторы (например, А. И. Орлов) считают, что данная группа вовсе не относится к кластеризации и противопоставляют её под названием «дискриминация», то есть выбор отнесения объектов к одной из известных групп (обучающих выборок):

1) K-средних;

2) К-медиан;

3) EM-алгоритм;

4) алгоритмы семейства FOREL;

5) дискриминантный анализ.

2. Подходы на основе систем искусственного интеллекта: весьма условная группа, так как методов очень много и методически они весьма различны:

1) метод нечеткой кластеризации C-средних (C-means);

2) нейронная сеть Кохонена;

3) генетический алгоритм.

4. Логический подход. Построение дендрограммы осуществляется с помощью дерева решений.

5. Теоретико-графовый подход:

1) графовые алгоритмы кластеризации.

6. Иерархический подход. Предполагается наличие вложенных групп (кластеров различного порядка). Алгоритмы в свою очередь подразделяются на агломеративные (объединительные) и дивизивные (разделяющие). По количеству признаков иногда выделяют монотетические и политетические методы классификации.

Иерархическая дивизивная кластеризация или таксономия. Задачи кластеризации рассматриваются в количественной таксономии.

7. Другие методы:

1) статистические алгоритмы кластеризации;

2) ансамбль кластеризаторов;

3) алгоритмы семейства KRAB;

4) алгоритм, основанный на методе просеивания;

5) DBSCAN и др.

**2 Общие достоинства и недостатки, действенность выбранного научного метода**

Для темы научных исследований “……………………………………………..” была выбрана наиболее подходящая разновидность кластерного анализа – EM-алгоритм (аргументировать).

EM-алгоритм (англ. Expectation-maximization (EM) algorithm) – алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных [ ]. Каждая итерация алгоритма состоит из двух шагов. На E-шаге (expectation) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На M-шаге (maximization) вычисляется оценка максимального правдоподобия, таким образом увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага на следующей итерации. Алгоритм выполняется до сходимости. Часто EM-алгоритм используют для разделения смеси гауссиан.

Для EM-алгоритма выполнен анализ достоинств, недостатков и действенность выбранного научного метода “Кластерный анализ” для темы научных исследований.

Достоинства EM-алгоритма:

а) …………………………………………………………………………………………………..;

б) …………………………………………………………………………………………………..;

………………………………………………………………………………………………………

Недостатки EM-алгоритма:

а) …………………………………………………………………………………………………..;

б) …………………………………………………………………………………………………..;

………………………………………………………………………………………………………

Действенность Е-алгоритма для темы научных исследований “…………………….” заключается в следующем:

а) …………………………………………………………………………………………………..;

б) …………………………………………………………………………………………………..;

………………………………………………………………………………………………………

**3 Выявление и сущность антипода метода кластерного анализа**

Суть метода кластеризации заключается в упорядочивании объектов в однородные группы на основе собранных данных об объектах.

Антиподом будет являться метод, который проводит обратное действие над однородными группами. Такого метода в чистом виде не существует. Так как в исследовании предполагается построение модели, то предлагается в качестве антипода методы снижения размерности, которые позволяют отбросить несущественные собранные данные об объекте. Это позволить упростить разрабатываемую модель.

В статистике, машинном обучении и теории информации снижение размерности – это преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных [ ]. Преобразование может быть разделено на отбор признаков и выделение признаков

Метод отбора признаков пытается найти подмножество исходных переменных (которые называются признаками или атрибутами). Есть три стратегии — стратегия фильтра (например, накопление признаков), стратегия обёртывания (например, поиск согласно точности) и стратегия вложения (выбираются признаки для добавления или удаления по мере построения модели, основанной на ошибках прогнозирования).

Проекция признаков преобразует данные из пространства высокой размерности к пространству малой размерности. Преобразование данных может быть линейным, как в методе главных компонент, но существует большое число техник нелинейного понижения размерности. Для многомерных данных может быть использовано тензорное представление для снижения размерности через полилинейное обучение подпространств.

Существуют следующие методы понижения размерности:

1) метод главных компонент (МГК);

2) неотрицательное матричное разложение (НМР);

3) ядерный метод главных компонент (ЯМГК);

4) основанный на графах ядерный МГК;

5) линейный дискриминантный анализ (ЛДА);

6) канонический корреляционный анализ;

7) обобщённый дискриминантный анализ (ОДА);

8) автокодировщик.

**2 Общие достоинства и недостатки, действенность антипода научного метода**

Для темы научных исследований “……………………………………………..” была выбрана наиболее подходящая разновидность антипода – метод главных компонент (аргументировать).

Метод главных компонент (англ. principal component analysis, PCA) – один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации [ ]. Изобретён К. Пирсоном в 1901 году. Применяется во многих областях: в эконометрике, биоинформатике, обработке изображений, для сжатия данных, в общественных науках.

Вычисление главных компонент может быть сведено к вычислению сингулярного разложения матрицы данных или к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных. Иногда метод главных компонент называют преобразованием Кархунена-Лоэва или преобразованием Хотеллинга (англ. Hotelling transform).

Для МГК выполнен анализ достоинств, недостатков и действенность выбранного научного метода “Снижение размерности” для темы научных исследований.

Достоинства МГК:

а) повышается быстродействие и снижается объем памяти;

б) применим к любым данным, а не только к нормально распределенным;

………………………………………………………………………………………………………

Недостатки МГК:

а) не всегда эффективно снижает размерность при заданных ограничениях на точность;

б) снижается эффективность при данных со сложной топологией;

………………………………………………………………………………………………………

Действенность МГК для темы научных исследований “………” заключается в следующем:

а) …………………………………………………………………………………………………..;

б) …………………………………………………………………………………………………..;

………………………………………………………………………………………………………

**Выводы**

1. Для темы научных исследований “ ” лучше всего подходит научный метод – “Кластерный анализ” потому что, (выразить в краткой форме)………………………………………

Было выявлено семь групп разновидностей кластерного анализа.

Для темы научных исследований наиболее подходит разновидность кластерного анализа: EM-алгоритм потому что, (выразить в краткой форме)…………………………………………………

2. Наиболее важным достоинством E-алгоритма является …………………………………….

В качестве наиболее существенного недостатка E-алгоритма можно отметить .…………………….

Наибольшую действенность при проведении исследований E-алгоритм будет оказывать на ………………………………………………………………….. ……………………………………….…

3. В качестве антипода кластерного анализа определены методы понижения размерности потому что, (выразить в краткой форме)…………………………………………………………………

Было установлено восемь групп методов понижения размерности.

Для темы научных исследований наиболее подходит метод понижения размерности: метод главных компонент потому что, (выразить в краткой форме)…………………………………………

4. Наиболее важным достоинством МГК является …………………………………….

В качестве наиболее существенного недостатка МГК можно определить……………………

Наибольшую действенность при проведении исследований МГК будет оказывать на (выразить в краткой форме)…………………………………………………………………………….…

PS.: а) желтый фон – заполнить требуемой информацией.

б) зеленый фон – рекомендации, которые необходимо удалить.