Секция «Модели и технологии цифровой трансформации экономики»

ДИВИЗИВНЫЙ МЕТОД АНАЛИЗА БОЛЬШИХ ДАННЫХ

Г.О. Абросимова
Научный руководитель А.И. Ильина
Самарский национальный исследовательский университет имени академика С.П. Королева

Большие данные – разнообразие данные, объем которых постоянно увеличивается.

Актуальность работы: В наше время развитие технологий дошло до уровня, когда анализировать только простые данные, является недостаточным. Необходимо анализировать многомерные большие данные, которые хранят в себе различную по своей природе информацию.

Научная новизна: Демонстрация алгоритма классификации без обучения многомерных данных размерностью 14x14.

Практическая значимость: Выявление пар различных объектов и структуризация в однородные по свойствам кластеры.

Цель исследования: выявление различных пар объектов и формирование структурно однородных по своим свойствам пар кластеров.

Задачи исследования:

11. Формирование исходного массива данных с использование весовой евклидовой метрики.
12. Формирование матрицы квадратов расстояний с использование евклидовой метрики.
13. Изучение и применение дивизимного метода.
14. Графическая визуализация результатов кластеризации.
15. Анализ структуры полученных кластеров.
16. Выбор оптимального варианта кластеризации.

$$(1) d_{ij} = \sqrt{\sum_{k=1}^{m} w_i \left(x_{ik} - x_{jk}\right)^2}$$

Суть иерархической дивизивной кластеризации: последовательное разделение больших кластеров на меньшие.

Алгоритм:

1. Все объекты принадлежат одному кластеру.
2. Находится d max.
3. Разделение кластера на меньшие.

4. Графическое фиксирование этапа кластеризации.

Таблица 3. Исходный массив данных.

№	1	2	3	4
1	1829,432	899,8865	0,04655	7,62642
2	307,5396	152,3593	0,0475	7,55595
3	359,022	193,7863	0,0399	7,68123
…	…	…	…	…
12	1437,443	788,9542	0,03515	7,81434
13	1102,356	550,9791	0,0475	7,61076
14	559,0808	288,1478	0,03515	7,69689

Таблица 4. Матрица квадратов расстояний.

№	1	2	3	…	12	13	14
1	0	2874952	2660682	…	165961,2	650376	1988015
2	2874952	0	4366,65	…	1681934	790630	81711,5
3	2660682	4366,65	0	…	1517216	680132	48927,6
…	…	…	…	...	…	…	…
12	165961	1681934	1517216	…	0	168916	1022327
13	650376	790630	680132	…	168915,6	0	364228
14	1988015	81711,5	48927,6	…	1022327	364228	0

Таблица 5. Матрица расстояний.

№	1	2	3	…	12	13	14
1	0	1695,57	1631,16	…	407,3834	806,459	1409,97
2	1695,57	0	66,0806	…	1296,894	889,174	285,852
3	1631,16	66,0806	0	…	1231,753	824,701	221,196
…	…	…	…	…	…	…	…
12	407,383	1296,89	1231,75	…	0	410,993	1011,1
13	806,459	889,174	824,701	…	410,9935	0	603,513
14	1409,97	285,852	221,196	…	1011,102	603,513	0

Таблица 6. Протокол кластеризации.

кластер	кластер	d(max)
1	2	1695,568
1	13	806,459
1	12	407,383
2	10	658,146
10	11	327,662
2	8	308,469
6	8	31,2706

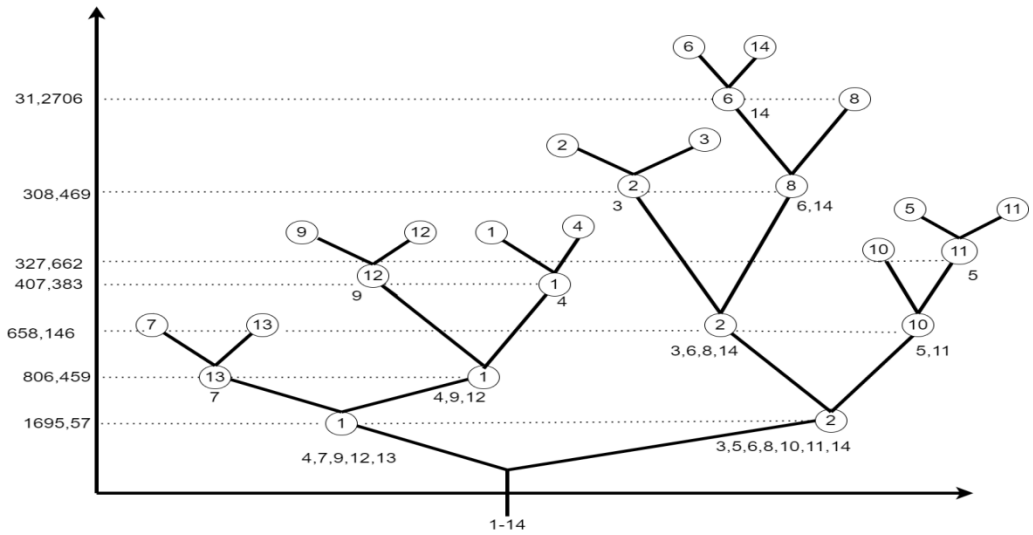

Рисунок 4. Графическая визуализация кластеризации.

Таблица 7.Пары объектов с минимальным диаметром расстояний.

Объекты	7 и 13	12 и 9	1 и 4	2 и 3	6 и 14	11 и 5
d_{ij}	82,90766	25,63547	78,33636	66,08063	7,501342	96,3254

Критерии качества:

$$(2) F_1 = \sum_{i=1}^{k} \sum_{i=1}^{p} d^2(x_i; \bar{x}_l)$$

$$(3) F_2 = \sum_{i=1}^{k} \sum_{j \in S}^{p} d_{ij}^2$$

$$(4) F_2 = \sum_{i=1}^{k} \sum_{i \in S_i} \sigma_{ij}^2$$

Варианты кластеризации:

1. S(1;4;7;9;12;13)
   S(2;3;5;6;8;10;11;14)
2. S(10;11;5)
   S(2;3;6;8;14)
   S(1;4;7;9;12;13)

Таблица 8.Выводы по расчетам критериев качества.

	F1	F2	F3
первый	855207,3	5713545	267215,7
второй	703608,6	3968661	161879,2
МИН	703608,6	3968661	161879,2

Оптимальный вариант кластеризации - второй

Выводы по работе:

1. Формирование исходного массива данных с использование евклидовой метрики.
2. Формирование матрицы квадратов расстояний и матрицы расстояний.
3. Изучение и применение дивизивных методов.

4. Графическая визуализация результатов кластеризации.
5. Выбор оптимального варианта кластеризации

Таким образом, мы поняли: Классификация без обучения позволяет получить качественно однородные группы кластеров.

Изучаемые показатели позволяют формировать стратегию развития объектов диаметрально расхожих.

Учитывая весовую кластеризацию с применением евклидовой метрики, происходит выявление доминирующих показателей выявляющих субъекты, которые не удовлетворяют условиям расхожести.

Список использованных источников

1. Анализ больших наборов данных Джеффри Дэвид Ульман 29 января 2022 г.
2. Теоретический минимум по Big Data. Всё что нужно знать о больших данных 18 сентября 2018 г.

МНОГОМЕРНЫЙ ПОДХОД ПРИ ВЫЯВЛЕНИИ РАЗЛИЧИЙ И СТРУКТУРИЗАЦИЯ С ОБУЧЕНИЕМ

Н.А. Базанов

Научный руководитель А.Ю. Трусова

Самарский национальный исследовательский университет имени академика С.П. Королева

Актуальность работы. Актуальность обусловлена изучением факторов, призванных повысить уровень инновационного развития в ПФО, используя многомерные статистические методы анализа. Показатели инноваций формируют стратегию развития общества и определяют дальнейшие перспективы. Инновации играют важную роль в развитии общества.

Научная новизна заключается в использовании многомерных подходов при анализе показателей влияющих на инновации.

Практическая значимость состоит в том, что инновации являются активным звеном всех сфер жизни общества и вызывают в них прогрессивные изменения. Преимуществами инноваций являются появление новых профессий, интеллектуализация условий труда, повышение уровня образованности и культуры, однако недостатками являются разрушение такого источника экономического роста, как полная занятость [1].

Цель работы: выявить различия и структурировать многомерные данные в сфере инноваций в ПФО в период с 2017 – 2020 годы.

Для достижения цели были выполнены следующие задачи:

1) сформировать выборку по показателям инноваций;

2) проверить гипотезы о равенстве многомерных показателей многомерному постоянному вектору (стандарту);