***Факультет инфокоммуникационных систем и технологий***

**КАФЕДРА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И УПРАВЛЯЮЩИХ СИСТЕМ**

**Статья**

**На тему «Кластерный анализ рынка компьютерных игр методом K-средних»**

**Выполнил:** *студент гр. ПМИ 19/2 Баранов О.В.*

Королев 2023

**Оглавление**

[Введение 3](#_Toc129260169)

[1. Описание предметной области 6](#_Toc129260170)

[2. Постановка математической задачи 7](#_Toc129260171)

[2.1. Минимального расстояния между кластерами 7](#_Toc129260172)

[2.2. Максимальное расстояние между кластерами 8](#_Toc129260173)

[2.3. Расстояние между центрами кластеров 9](#_Toc129260174)

[3. Алгоритм выполнения задания курсового проекта 10](#_Toc129260175)

[3.1. Среднее расстояние между кластерами 10](#_Toc129260176)

[3.2. Минимальное расстояние между кластерами 16](#_Toc129260177)

[3.3. Максимальное расстояние между кластерами 17](#_Toc129260178)

[Заключение 19](#_Toc129260179)

[Список используемых источников 20](#_Toc129260180)

# **Введение**

Кластерный анализ — 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 [1]. Темой работы является кластерный анализ рынка компьютерных игр методом K-средних.

Для кластеризации были взяты «шутеры» (Стрелялка, англ. shooter — «стрелок» [2]):

1. Counter-Strike 1.6;
2. Counter-Strike: Source;
3. Counter-Strike: Global Offensive;
4. Tom Clancy’s Rainbow Six: Siege;
5. Call of Duty: Modern Warfare;
6. Call of Duty: Modern Warfare 2 (2009);
7. Call of Duty: Modern Warfare 3;
8. Call of Duty: Black Ops;
9. Call of Duty: Black Ops II;
10. Call of Duty: Black Ops III;
11. Call of Duty: Black Ops 4;
12. Call of Duty: Ghosts;
13. Call of Duty: Advanced Warfare;
14. Call of Duty: World at War;
15. Call of Duty: Modern Warfare II (2022);
16. Battlefield 3;
17. Battlefield 4;
18. Battlefield Hardline;
19. Battlefield 1;
20. Battlefield V;
21. Battlefield 2042;
22. Warface;
23. Payday: The Heist;
24. Payday 2;
25. Team Fortress Classic;
26. Team Fortress 2;
27. Overwatch;
28. Overwatch 2;
29. Quake Champions;
30. Star Wars: Battlefront.

Также каждая игра имеет 30 схожих объектов выборки:

1. количество карт,
2. количество вариантов оружия,
3. количество видов оружия,
4. количество видов транспорта,
5. количество классов персонажей,
6. количество видов дополнительного снаряжения,
7. количество режимов,
8. количество дополнений,
9. количество рангов/званий,
10. количество сюжетных компаний.

Задачами работы является:

1. заполнить таблицу данных, на основании которой и будет осуществляться кластеризация;
2. все показатели из таблицы исходных данных необходимо нормировать – привести к диапазону от 0 до 1;
3. выполнить кластерный анализ множества объектов методом K-средних;
4. вычислить минимальное, максимальное и среднее расстояния между центрами кластеров;
5. построить для каждого кластера гистограмму расстояний объектов от центра кластера.

# **Описание предметной области**

Наряду с иерархическими методами классификации существует многочисленная группа так называемых итеративных методов кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров. Например, выбранное случайным образом число кластеров может не только увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок, а затем уже подбирать начальное разбиение и статистический критерий для работы итерационного алгоритма. Как и в иерархическом кластерном анализе, в итерационных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров. Но для окончательного решения вопроса о структуре изучаемой совокупности можно испробовать несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры. Тогда появляется возможность выбрать наилучшее разбиение по задаваемому критерию качества [3].

Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга [4].

# **Постановка математической задачи**

Пусть имеется n объектов, каждый из которых характеризуется p признаками x1, x2 , …, xn. Эти наблюдения необходимо разбить на k кластеров. Сначала из значений исследуемой совокупности отбираются случайным образом или задаются из каких-либо априорных соображений. После чего осуществляется построение различный комбинаций, среди которых выбирается один, который является оптимальным. Следующим этапом является вычисления расстояния между кластерами. Для расчёта использовалось евклидово расстояние.

(2.1)

$$d=\sqrt{\sum\_{i}^{}(x\_{Ai}-x\_{Bi})^{2}}$$

# **Минимального расстояния между кластерами**

Чтобы найти минимальное расстояние между кластерами необходимо построить таблицу расстояний между объектами, которое рассчитается по формуле 2.1 (рис 1).



Рисунок 1 –таблица расстояний между объектами кластеров

После нахождения минимального расстояния между кластерами, которое на рисунке 1, равно 2,5 условных единиц. Следующим этапом отмечаются координаты точек A2 и B2, которые образуют прямую, которая и является минимальным расстоянием между кластерами (рис 2).



Рисунок 2 – минимальное расстояние между кластерами A и B

# **Максимальное расстояние между кластерами**

Чтобы найти максимальное расстояние между кластерами необходимо также, как и для нахождения минимального расстояния между кластерами, построить таблицу расстояний между объектами, которое рассчитается по формуле 2.1 (рис 3).



Рисунок 3 –таблица расстояний между объектами кластеров

После нахождения максимального расстояния между кластерами, которое на рисунке 3, равно 6,403 условных единиц. Следующим этапом отмечаются координаты точек A4 и B1, которые образуют прямую, которая и является максимальным расстоянием между кластерами (рис 4).



Рисунок 2 – максимальное расстояние между кластерами A и B

# **Расстояние между центрами кластеров**

Чтобы найти расстояние между центрами кластеров необходимо найти среднее значение каждого показателя кластеров.

(2.2)

$\frac{\sum\_{i=1}^{}x\_{ji}}{n}$,

где j – номер кластера, а n – кол-во элементов вида i.

Так с исходными данными (рис 5) получились расстояние равное 4,53 условных единиц (рис 6).



Рисунок 5 – таблица с данными по кластерам A и B



Рисунок 5 – расстояние между центрами кластеров A и B

# **Алгоритм выполнения задания курсового проекта**

# **Среднее расстояние между кластерами**

Для разделения объектов по кластерам необходимо начать с заполнения таблицы с данными, которая включает в себя 30 объектов (игр) и 10 количественных показателей, которые характерны для всех игр (рис 6).



Рисунок 6 – таблица с данными для кластерного анализа

Для удобства названия количественных показателей были заменены на условные переменные (рис 7).



Рисунок 7 – обозначения переменных

После заполнения таблицы данными необходимо совершить нормировку данных. Первым делом нужно найти минимум и максимум по каждому показателю, а потом найти L, где i – это какая-либо из условных переменных.

(2.3)

$$L\_{i}=max\_{i}-min\_{i}⁡$$

Далее заполняем таблицу с нормированными значениямиyj. Для примера того, как нормируются данные в формуле 2.4 будет показано как нормировался показатель x1.

(2.4)

$$y\_{x\_{1}}=\frac{x\_{1}-min\_{x}}{L\_{x}}$$

Все остальные данные считаются по аналогии уравнения 2.4 (рис 8).



Рисунок 8 – таблица нормированных показателей

Следующим этапом является выявление центроидов каждого из кластеров. Их было взято 3, так как на большее количество разбивать не получилось, так как после преобразований оставалось всего 3 группы. На втором и третьем листах работы в Excel были показаны примеры того, как осуществлялся выбор центроидов, а на четвертом листе были найдены оптимальные центроиды.

Работа совершенная на втором – четвертом листах в среде Excel не отличается друг от друга по использованию формул и логика на них абсолютно одинаковая, поэтому объяснение выполнения кластеризации методом K-средних будет показана только на оптимальных центройдах.

 В качестве центроидов были взяты 1, 21 и 30 элементы выборки. После выполнялось нахождение точек, которые относятся к конкретному центроиду. Строится таблица, где по строкам пишется номер элемента выборки, а по столбцам пишется номер центроида, каждый элемент считается по формуле 2.1. Пусть элементы таблицы будут обозначаться переменной rij, где i – номер центроида, а j номер элемента выборки.

(2.5)

$$r\_{ij}=\sqrt{\sum\_{i=1}^{n}(y\_{i\_{1}}-y\_{ц1})^{2}}$$

После построения таблицы находятся минимальные расстояния центроидов, а также отмечается к какому центроиду относятся показатели объекта (рис 9).



Рисунок 8 – таблица расстояний для распределение объектов на группы

После необходимо начертить график, который показывает разделение объектов на группы, а также отразить расстояние между центрами кластеров (рис 9), а также построить гистограмму отражающую значения центроидов (рис 10).



Рисунок 9 – расстояние от центров кластеров



Рисунок 10 – гистограмма кластеров

Одним из важных показателей является S, который является суммой минимумов расстояний между кластерами. Если показатель S при повторении действий уменьшается, то следует повторить тот же самый алгоритм, описанный выше. В данном случае он равняется 17,124 условных единиц. При повторных операциях его минимальное значение, которое получилось вывести составляло 15,088 условных единиц. Больше пытаться уменьшить значение S не следует, так как все показатели объединяться в один кластер. Результатом стал преобразований стал график (рис 11) и гистограмма (рис 12).



Рисунок 11 – финальное расстояние от центров кластеров



Рисунок 12 – финальная гистограмма кластеров

# **Минимальное расстояние между кластерами**

Для нахождения минимального расстояния необходимо построить таблицу с координатами кластеров (рис 13).



Рисунок 13 – таблица координат кластеров

Чтобы найти минимальное расстояние между кластерами необходимо построить таблицу расстояний между объектами, которое рассчитается по формуле 2.1 (рис 14).



Рисунок 14 – таблица расстояний между объектами кластеров

Минимальным расстоянием между кластерами равно 0 условных единиц (рис 15).



Рисунок 15 – минимальное расстояние от центров кластеров

# **Максимальное расстояние между кластерами**

Для нахождения максимального расстояния можно воспользоваться таблицей с координатами кластеров для нахождения минимального расстояния между кластерами (рис 13).

Чтобы найти максимальное расстояние между кластерами необходимо построить таблицу расстояний между объектами, которое рассчитается по формуле 2.1 (рис 16).



Рисунок 16 – таблица расстояний между объектами кластеров

Максимальное расстояние между кластерами равно 1 условной единиц (рис 17).



Рисунок 17 – минимальное расстояние от центров кластеров

## **Заключение**

В данной работе был сделан кластерный анализ методом K-средних, результатом которой являются графики минимального, максимального и среднего расстояний между кластерами, а также построена гистограмма расстояний между центрами кластеров.

В работе были выполнены следующие задачи:

1. выполнен кластерный анализ множества объектов методом K-средних;
2. вычислено минимальное, максимальное и среднее расстояния между центрами кластеров;
3. отображено графически выборка и динамика оценивания координат центров кластеров;
4. построена для каждого кластера гистограмму расстояний объектов от центра кластеров.

# **Список используемых источников**

1. Кластерный [Электронный ресурс]. Сайт: <https://ru.wikipedia.org>. URL: [https://ru.wikipedia.org/wiki/Кластерный\_анализ](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7) (дата обращения 10.12.2022)
2. Шутер [Электронный ресурс]. Сайт: <https://ru.wikipedia.org>. URL: [https://ru.wikipedia.org/wiki/Шутер](https://ru.wikipedia.org/wiki/%D0%A8%D1%83%D1%82%D0%B5%D1%80) (дата обращения 10.12.2022)
3. Метод к-средних [Электронный ресурс]. Сайт: <https://studfile.net>. URL: [https://studfile.net/preview/1938850/page:11/](https://studfile.net/preview/1938850/page%3A11/) (дата обращения 10.12.2022)
4. Тема 9. Кластерный анализ [Электронный ресурс]. URL: <https://nafi.ru/upload/spss/Lection_9.pdf> (дата обращения 10.12.2022)
5. Математика для всех «Класстерный анализ. Иерархическая кластеризация. Метод k-средних.» [Видеозапись]. Сайт: [www.youtube.com](http://www.youtube.com). URL: <https://www.youtube.com/watch?v=GOcFenttYIw>