МИНИСТЕРСТВО СПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский университет спорта «ГЦОЛИФК»

**Кафедра теории и методики компьютерного спорта и прикладных**

**компьютерных технологий**

Ермаков А.В.

**Кластерный анализ в исследовании спортивной деятельности**

для обучающихся

по направлению подготовки 49.04.03 Спорт

г. Москва, 2024 г.

Утвержден и рекомендован

к использованию в образовательном процессе

Экспертно-методическим советом

Протокол № от « » 2025г.

Ермаков А.В.

Построение регрессионных моделей для исследования спортивной деятельности в единоборствах: учебное пособие для обучающихся по направлению подготовки 49.04.03 Спорт. - г. Москва: РУС «ГЦОЛИФК», 2024. - 48 с.

Рецензент:.

Учебное пособие посвящено построению регрессионных моделей для исследования спортивной деятельности в единоборствах и адресован обучающимся по программам бакалавриата для подготовки к практическим занятиям и к зачету по дисциплине Б1.В.10 Искусственный интеллект и машинное обучение в спорте

Содержание

Оглавление

[Введение 4](#_Toc189570310)

[Теоретико-методические основы регрессионного анализа спортивной деятельности 7](#_Toc189570311)

[Использование алгоритмов кластерного анализа K-means и DBSCAN для проведения исследований спортивной подготовки в компьютерном спорте. 14](#_Toc189570312)

[Список литературы 27](#_Toc189570313)

# Введение

Целью данного учебного пособия является подготовка магистрантов, обучающихся по направлению подготовки 49.04.03 Спорт к решению исследовательских задач в сфере физической культуры и спорта с использованием средств, и методов искусственного интеллекта и машинного обучения для автоматизации и оптимизации исследовательских задач, снабжение их теоретическими знаниями и практическими умениями анализа данных. В результате применения пособия будущие магистры обучаются эффективно использовать методы кластерного анализа для исследования спортивной деятельности при обработке данных различного размера. Обучаемые по программе магистратуры по дисциплине «Б1.В.10 Искусственный интеллект и машинное обучение в спорте» получают набор знаний и умений согласно следующим компетенциям (табл.1)

Таблица 1. – Компетенции, реализуемые при использовании магистрантами методов кластерного анализа

|  |  |
| --- | --- |
| УК-1.2. | Знает основные аспекты системного подхода как базы научного осмысления интегративной сущности физической культуры и спорта. |
| УК-1.3. | Умеет критически анализировать, оценивать научно-методический и исследовательский опыт в аспекте выявления актуальных научных проблем, изучения закономерностей развития процессов научного знания в избранном виде профессиональной деятельности. |
| ПК-3.1 | Знает современные цифровые технологии, используемые в прикладных научных исследованиях. |

Использование цифровых технологий для анализа данных являются одним из актуальных средств спортивной педагогической науки[[1]](#footnote-1)[[2]](#footnote-2)[[3]](#footnote-3). В условиях значительно увеличившегося потока данных очень важно использовать специализированные методы обработки данных в том числе и большого размера[[4]](#footnote-4). В ходе управления процессом спортивной подготовки и анализе соревновательной деятельности возможно произвести разделение данных на кластеры для последующего анализа, это позволит повысить качество производимого анализа в целях повышения объективности принятия решений улучшения качества проводимого анализа[[5]](#footnote-5)[[6]](#footnote-6). Сегодня кластерный анализ широко применяется в исследованиях спортивной деятельности за рубежом[[7]](#footnote-7). В нашей стране использование кластерного анализа в изучении спортивной деятельности также является ценным инструментом позволяющим анализировать в том числе и данные большого размера[[8]](#footnote-8).

Поэтому данное пособие может быть использовано во всех практических и исследовательских задачах, в которых магистрантам сложно обнаружить явные закономерности в исследуемых данных и одним из решений является разделение данных на кластеры.

# Теоретико-методические основы регрессионного анализа спортивной деятельности

Кластерный анализ представляет собой задачу группировки множества объектов для получения определённого количества подмножеств таким образом, чтобы объекты в одной группе (называемой кластером) были похожи согласно условиям, определённым для анализа, друг на друга больше, чем на объекты в других группах. (кластерах). Такой подход используют для статистического анализа, для компьютерного зрения при распознавании и анализе изображений, сжатия данных, компьютерной графики и машинного обучения. Он обычно направлен на изучение внутренней структуры сложного набора данных, который не может быть описан только с помощью классической статистики второго порядка (таких как выборочное среднее и ковариация). В спортивной подготовке существуют естественные признаки для разделения генеральной совокупности на группы по полу, возрасту, весу, спортивной квалификации и тому подобное. Однако, в ходе исследования возможна ситуация, когда исследователь не может выявить закономерности в изучаемых данных и в этом случае наиболее оправдано проведение кластерного анализа, который может позволить выдвинуть гипотезу о подобной закономерности. Иными словами, кластерный анализ может быть назван неконтролируемой классификацией в отличии от обычной классификации предполагающей, что классы уже известны и исследователю нужно только отнести экземпляр множества к одному или другому классу.

Кластерный анализ не ограничен одним алгоритмом, а представляет из себя целое семейство алгоритмов и задач. Главное отличие алгоритмов, заключается в понимании того, что представляет из себя кластер и как более эффективно можно его найти. Наиболее часто используемые определения понятия кластеров включают в себя определения использующие наименьшие расстояния между членами кластера, плотность конкретной области пространства данных, а также интервалы или определенные статистические распределения. Таким образом, кластерный анализ можно свести к задаче многокритериальной оптимизации. Выбор наиболее подходящего алгоритма и настройки его гиперпараметров (включая такие параметры, как используемая функция расстояния между объектами, порог плотности множества или количество ожидаемых кластеров) зависят от особенностей анализируемого набора данных и предполагаемого использования результатов, что кроме математической корректности предъявляет повышенные требования к знанию предметной области. Кластерный анализ как таковой не выполняется механистически, а представляет собой итеративный процесс открытия знаний, включающий пробы и их оценки. Часто приходится многократно изменять гиперпараметры модели, пока результат не достигнет требуемых свойств.

Считается, что первая работа использовавшая кластерный анализ была в области антропологии в 1932 году[[9]](#footnote-9) далее его использовали в психологии в 1938 году[[10]](#footnote-10) после этого уже в сороковых годах стал широко использоваться ведущими исследователями например Кеттелл использовал кластерный анализ для классификации теории черт в психологии личности[[11]](#footnote-11).

Понятие «кластера» не может быть точно определена, что является одной из причин, почему существует так много алгоритмов кластеризации[[12]](#footnote-12).

Кластерный анализ является одним из основных методов в интеллектуальном анализе данных[[13]](#footnote-13). Традиционно для любого исследовательского пути существует два основных методологических подхода дедуктивный и индуктивный. Один подход рассматривает кластеризацию как сегментацию неоднородного множества на ряд более однородных подмножеств[[14]](#footnote-14). Другой определяет кластеризацию как поиск подмножеств в наборе данных «по некоторому естественному критерию сходства»[[15]](#footnote-15). При этом основной фундаментальный вопрос заключается в том, находятся ли два элемента множества в одном кластере или нет. Определение естественного критерия сходства элементов является основанием к определению того, как учитывать сходство между группами. Некоторое интуитивное представление о том, что составляет достаточно плотные группы, приводит к использованию принципа индукции. Фактически, большинству авторов трудно описать кластеризацию без некоторого предложения критериев группировки. Например, «объекты разделяются на кластеры или группируются на основе принципа максимизации межклассового сходства и минимизации внутриклассового сходства»[[16]](#footnote-16).

Уже с 1967 года ситуация несколько изменилась и было заявлено, что кластерный анализ рассматривается скорее как помощь исследователям в получении качественного и количественного понимания большого количества многомерных данных, чем только вычислительный процесс, который находит некоторую уникальную и окончательную группировку для данных[[17]](#footnote-17). Позже, из-за своей неконтролируемой, описательной и обобщающей природы, кластерный анализ данных также стал основным методом анализа данных и обнаружения новых знаний. Непрерывно растущее число больших многомерных наборов данных сделало актуальным и ускорило разработку новых алгоритмов кластерного анализа[[18]](#footnote-18). Вследствие того, что особенно важным кластерный анализ становится при анализе данных большого размера, возникает необходимость и в разнообразии подходов для решения этой задачи. Структура подходов к кластерному анализу больших данных представлена на рисунке 1[[19]](#footnote-19).

В случае данных большого размера согласно схеме, на рисунке 1 выделяют группу из нескольких алгоритмов.

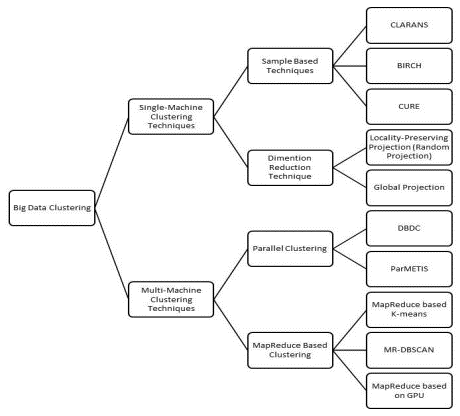


Рис.1 - Структура подходов к кластерному анализу больших данных

1. Методы кластеризации на одной машине (Single-machine clustering techniques). В них входят:

1.1. Методы, основанные на выборке (Sampling Based Techniques), которые были первыми попытками улучшить скорость и масштабируемость работы алгоритмов, в основном они основывались на экспоненциальном поиске. Эти алгоритмы называются алгоритмами, основанными на выборке, потому что вместо выполнения кластеризации на всем наборе данных, они выполняют алгоритмы кластеризации на выборке наборов данных, а затем обобщают ее на весь набор данных. Это ускоряет алгоритм, поскольку вычисления выполняются для меньшего количества данных, и, следовательно, сложность и пространство памяти, необходимое для процесса, уменьшаются.

1.2 Методы сокращения размерности (Dimension Reduction Techniques). Хотя сложность и скорость алгоритмов кластеризации связаны с количеством наблюдений в наборе данных, но с другой стороны количество параметров также является значимым аспектом. Фактически, чем больше параметров имеют данные, тем выше сложность и это означает большее время выполнения. Методы выборки могут использоваться для уменьшения размерности набора данных за счёт уменьшения количества учитываемых параметров.

2 Методы кластеризации на нескольких машинах (Multi-Machine clustering techniques).

Хотя методы выборки и сокращения размерности, используемые в алгоритмах кластеризации на одной машине, улучшают масштабируемость и скорость алгоритмов, но в настоящее время рост размера данных происходит намного быстрее, чем увеличение объёмов памяти и мощности процессоров, следовательно, одна машина с одним процессором и памятью не может обрабатывать терабайты и петабайты данных, и это подчеркивает необходимость алгоритмов, которые могут быть запущены на нескольких машинах. Этот подход позволяет разбить большой объем данных на более малые части, которые можно загрузить на разные машины. Алгоритмы кластеризации на нескольких машинах делятся на две основные категории:

* Неавтоматизированное распределение – параллельное
* Автоматизированное распределение – MapReduce

2.1 Параллельная кластеризация

Хотя параллельные алгоритмы добавляют сложности распределения, но это может быть оправдано из-за достигаемых улучшений в масштабировании и скорости алгоритмов кластеризации. В параллельной кластеризации разработчики занимаются не только проблемами самой кластеризации, но и деталями процесса распределения данных между различными машинами, доступными в сети, что делает его достаточно сложным и трудоемким. Разница между параллельными алгоритмами и MapReduce заключается в том, что MapReduce обеспечивает техническим специалистам удобство и избавляет их от ненужных сетевых проблем, таких как распределение данных, отказоустойчивость и т. д., обрабатывая их автоматически.

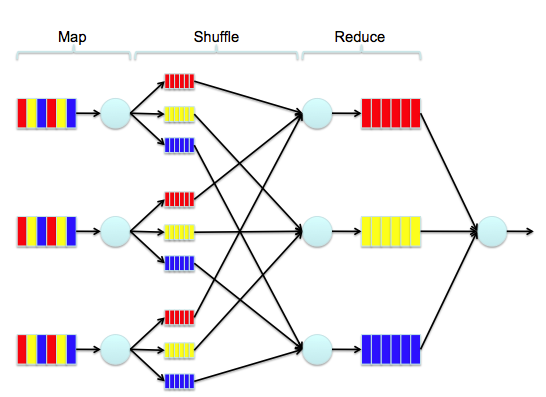


Рис.2 – Принципиальная схема работы фреймворка MapReduce

2.2 Метод MapReduce

Примером MapReduce служит фреймворк, который был представлен Google, и Hadoop как его версия с открытым исходным кодом. Суть MapReduce состоит в разделении информационного массива на части, параллельной обработке каждой части на отдельном узле и финального объединения всех результатов (рис.2).

Работа MapReduce состоит из трёх основных шагов:

Map-шаг схож с одноимённой функцией Map функционального программирования. На данном шаге происходит предварительная обработка, разделение получаемого для обработки списка на части, передачи рабочим узлам, которые применяют функцию Map и осуществляют запись выходных данных в виде ключей во временном хранилище.

На Shuffle-шаге рабочие узлы перераспределяются данные на основе ключей, созданных на Map-шаге, так что все данные, принадлежащие к одному ключу, расположены на одном рабочем узле. Для этого мастер-узел получает входные данные задачи, разделяет их на части и передаёт другим рабочим узлам для предварительной обработки.

На Reduce-шаге происходит свёртка предварительно обработанных данных. Мастер-узел получает ответы от рабочих узлов и на их основе происходит решение задачи, которая изначально формулировалась.

Вне зависимости от величины набора данных любой кластерный анализ, проведённый алгоритмом или даже просто человеком, является гипотезой, предполагающей (или объясняющей) определённую группировку исследуемых данных и корректная интерпретация результатов зависит качества знаний в предметной области у исследователя.

**Вопросы для самоконтроля.**

Сущность и основание для проведения кластерного анализа

Многозначность определения понятия кластер

Методы, основанные на выборке

Методы сокращения размерности

Параллельная кластеризация

Метод MapReduce

# Использование алгоритмов кластерного анализа K-means и DBSCAN для проведения исследований спортивной подготовки в компьютерном спорте.

В качестве примера использования кластерного анализа в исследовании спортивной подготовки используем достаточно простой, интуитивно понятный и популярный алгоритм k-means. Для анализа был использован набор данных об игровой активности во время соревнований по популярной киберспортивной игре League of Legends. Данные включали в себя оценки по 40 признакам 9879 соревновательных матчей с высоким рейтингом (от Diamond I до Master). Таким образом количество данных достаточно велико для спортивной науки, но мал для необходимости разделять его для обработки на несколько машин. Однако может быть несколько сложно выделить определённые закономерности в данных и потому вполне оправдано использовать кластерный анализ для более глубокого их понимания и возможно для дальнейшего изучения отдельных кластеров.

Первый выбранный для анализа алгоритм k-means (k-средних) относится к методам машинного обучения «без учителя» (Unsupervised Learning), он достаточно простой, но как следствие не может похвастаться высокой точностью. Однако для получения новых знаний при исследовании спортивной подготовки его возможностей вполне достаточно и применение оправдано. Алгоритм разбивает множество элементов векторного пространства на заранее известное число кластеров, отсюда обозначение этого количества как k. Действие алгоритма заключается в стремлении минимизировать среднеквадратичное отклонение элементов каждого кластера от его центра масс (центроида) или говоря иначе минимизировать внутрикластерную дисперсию, одновременно максимизируя межкластерную дисперсию. Основная идея заключается в том, что на каждой итерации заново вычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центроидов оказался ближе по выбранной метрике. Итерации происходят до тех пор, пока вновь вычисленные центры масс не будут отличаться от ранее определённых сами кластеры при пересчёте остаются неизменными.

Метод k-средних пытается сгруппировать похожие элементы в три этапа:

* Выбирается оптимальное количество кластеров, это значение k из названия алгоритма
* Случайным образом инициализируются центроиды (или разделительные линии)
* Выбирается кластер, к которому принадлежит объект по признаку наибольшей близости к центроиду кластера относительно других центроидов.
* Помещение центроидов в соответствующие центры масс кластеров и пересчёт принадлежности объектов к кластерам по признаку наибольшей близости

Завершение работы алгоритма происходит, когда на какой-то итерации не происходит изменения положений центроидов и состава самих кластеров. Упорядочивая свойства алгоритма можно обозначить его достоинства и недостатки, влияющие на выбор исследователя следующим образом.

**К преимуществам алгоритма k-means относят:**

* Простота его реализации
* Масштабируемость алгоритма включая данные больших размеров
* Быстрое обучение для новых данных
* Обработка данных сложных форм и размеров.
* Универсальность для различных предметных областей

**К недостаткам алгоритма k-means относят:**

* Чувствительность к выбросам и сложным формам кластеров
* Исследователю необходимо заранее определить количество кластеров, на которые будут разделены исследуемые данные.
* Алгоритм очень чувствителен к случайному выбору начальных центров масс определяемых кластеров.
* Алгоритм не справляется с задачей, когда объект принадлежит к разным кластерам в равной степени (отстоит на одинаковое расстояние от центров масс разных кластеров) или не принадлежит ни одному.

В первую очередь можно использовать естественный критерий для разделения на кластеры в спортивной деятельности – это спортивный результат. В исследуемом случае с киберспортивными матчами по League of Legends можно рассмотреть деление на два естественных кластера – победа и поражение в матче.

Можно рассмотреть разделение данных на два кластера на примере визуализации двух параметров – собственных потерь и потерь соперника (рис.3).

В результате опоры только на два параметра можно увидеть действительно логичное разделение на два кластера к одному относятся команды с отношением уничтоженных юнитов соперника к собственным потерям большим единицы, а к другому, с коэффициентом соответственно от единицы меньше.

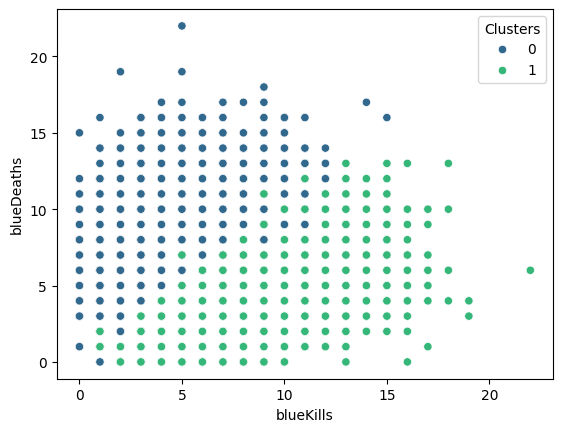


Рис. 3 – Визуализация разделения данных алгоритмом k-means на два кластера на примере двух параметров

При этом угол наклона говорит о том, что в данных матчах к победе приводила чуть более агрессивная тактика. Однако такой естественный подход далеко не единственный и зачастую имеет смысл определять оптимальное количество кластеров математическими методами. Одним из простых и достаточно наглядных методов определения оптимального количества кластеров является метод «локтя» (рис.4).

Суть метода заключается в измерении значения WCSS (within-cluster sum of squares) что представляет собой сумму квадратов отклонений объектов от центра соответствующих кластеров для разбиения на различное количество кластеров.

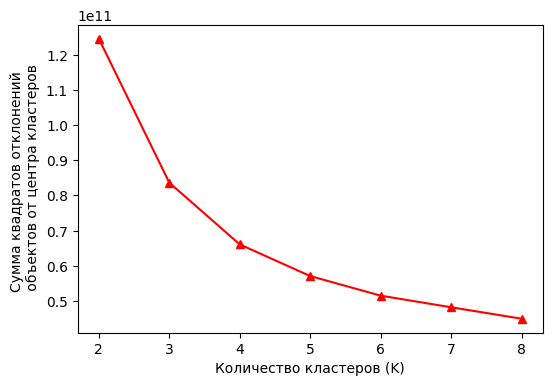


Рис.4 – Определение количества кластеров по методу «локтя»

В результате можно заметить убывание этого расстояния, то есть объекты в кластерах становятся всё ближе к центрам масс при увеличении количества кластеров в плоть до того, что каждый отдельный экземпляр данных будет относится к отдельному кластеру и расстояние в этом случае будет равно нулю. Но с увеличением количеств кластеров интерпретируемость построенной модели может значительно ухудшаться. Поэтому необходимо определить то оптимальное число кластеров, которое с одной стороны делает деление однозначным (внутрикластерные расстояния становятся малыми), а с другой стороны количество кластеров позволяет легко интерпретировать модель. Вид графика напоминает согнутую руку человека откуда и возникло название и место локтя, то есть перегиб на графике отражает то самое положение, когда увеличение количества кластеров не вносит значительный вклад в уменьшение внтурикластерного расстояния. Конечно оценка по методу локтя достаточно субъективна и зачастую возникает несколько кандидатов на точку перегиба и в нашем случае такой точкой могут быть соответствующие трём, четырём и даже пяти кластерам. Поэтому необходимо провести оценку для большего чем два количества кластеров. Пример для разделения на три кластера по прежним двум параметрам представлено на рисунке 5.

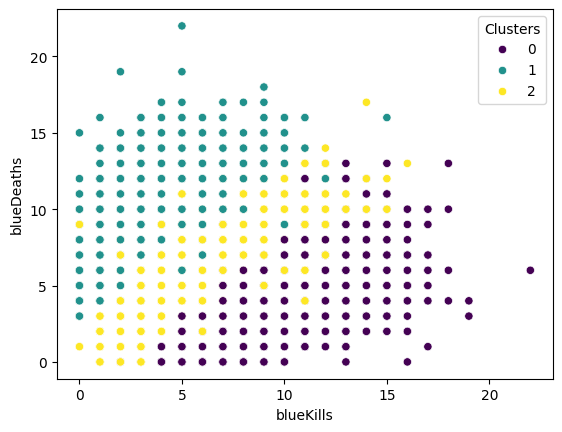


Рис. 5 – Визуализация разделения данных алгоритмом k-means на три кластера на примере двух параметров

Можно обратить внимание, что и в случае трёх кластеров можно интерпретировать кластеры как в первую очередь уничтожающие юниты соперника, теряющие свои и сбалансированный в котором существует паритет между собственными потерями и уничтожением юнитов соперника. Проверяя найденную тенденцию можно продолжить увеличение количества кластеров. Пример для разделения на четыре кластера по прежним двум параметрам представлено на рисунке 6.

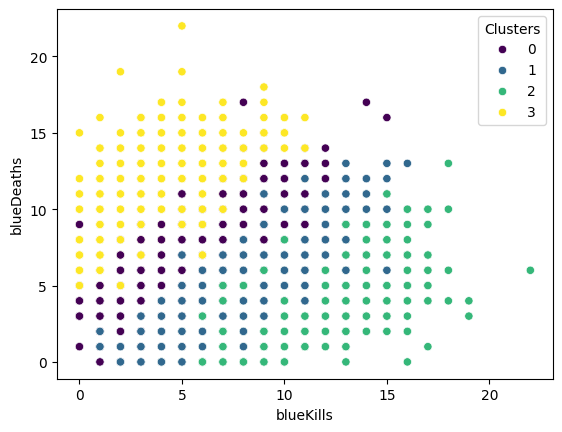


Рис. 6 – Визуализация разделения данных алгоритмом k-means на четыре кластера на примере двух параметров

Как видно на графике при дальнейшем увеличении количества кластеров возникает ситуация, когда, опираясь только на эти два параметра интерпретировать объектные закономерности отнесения конкретного объекта к одному или другому кластеру уже становится сложно и дальнейшее увеличение количества кластеров при опоре на эти два параметра не имеет смысла.

Для сравнения используем алгоритм, работающий по другим принципам. Для него не требуется заранее определять количество кластеров. Основными гиперпараметрами будут радиус окрестности и минимальное количество объектов, которое может находиться в одном кластере. Алгоритм имеет название DBSCAN (Density-Based Spatial Clustering of Applications with Noise) и исходя из названия он основывается прежде всего на значениях плотности объектов в кластере. Алгоритм выделяет области высокой концентрации объектов, которые разделяются областями с низкой плотностью. Объекты, которые находятся в области с низкой плотностью (ближайшие соседи которых лежат далеко друг от друга), определяются алгоритмом в качестве выбросов.

Более подробное описание работы алгоритма DBSCAN следующее:

1. Алгоритм берёт объект из набора данных и строит вокруг него буфер указанного исследователем радиуса. Если в буфер попадает количество объектов больше, чем минимально определённое количество, то данный объект становится корневым и вокруг него строится новый кластер. Если два корневых элемента находятся рядом, то они объединяются в один кластер. Таким образом проверяются все объекты исследуемого множества и в одном кластере может быть большое количество корневых объектов, но е меньше одного.
2. Далее алгоритм находит объекты, у которых в буфере меньше заданного количества соседей, но есть хотя бы один корневой. Подобные объекты становятся пограничными.
3. После выделения корневых и пограничных объектов остаются такие, в буфере которых находится меньше указанного числа соседей и нет корневых объектов. Именно такие объекты и будут считаться выбросами.

**К преимуществам алгоритма DBSCAN относят:**

1. Отсутствие требований заранее указывать количество кластеров;
2. Устойчивость к выбросам;
3. Способность находить кластеры сложной формы.

**К недостаткам алгоритма DBSCAN относят:**

1. Слабую эффективность при обработке кластеров разной плотности;
2. Необходимость выделения большого объём памяти для хранения расстояний между всеми объектами исследуемого набора данных;
3. Высокую чувствительность к выбору гиперпараметров, что по большому счёту аналогично выбору количества кластеров в алгоритме k-means. Часто применение алгоритма DBSCAN может сводиться к подбору гиперпараметров для получения удовлетворительного результата.
4. Вычислительную сложность алгоритма, которая может стать проблемой для больших наборов данных или для задач, где требуется быстрый анализ в реальном времени.
5. Малую эффективность алгоритма при анализе данных большой размерности.

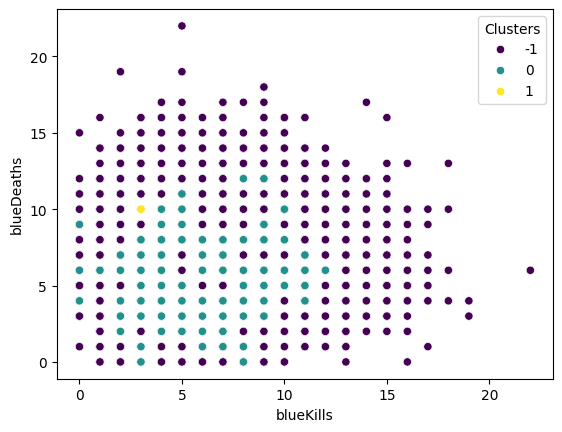


Рис. 7 – Визуализация разделения данных алгоритмом DBSCAN на три кластера на примере двух параметров

Подобрав гиперпараметры так, чтобы получилось три кластера можем также посмотреть визуализацию по выбранным ранее двум параметрам на рисунке 7. Плотность показывает в данном случае наиболее часто встречающиеся значения, объединённые в один кластер, что позволит выделить тот самый «средний матч» в дальнейшем анализируя который можно добиваться более устойчивых моделей.

Оценить результативность проведённого анализа можно используя большое количество различных метрик комбинируя их в зависимости от особенностей набора данных и требований предметной области. Теперь же рассмотрим комплекс метрик, помогающих оценить работу модели.

Коэффициент «Silhouette» или анализ кластерных силуэтов вычисляется с помощью среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера по каждому экземпляру. Также этот коэффициент определяют, как меру того, насколько объект похож на свой кластер по сравнению с другими кластерами можно вычислить среднее значение силуэта кластера по всем экземплярам и использовать его как метрику для оценки количества кластеров. метод графического представления результатов кластеризации, с помощью которого можно визуально оценить качество построенной кластерной модели. По форме силуэтов можно оценить качество кластерного анализа. Чем форма силуэтов ближе к прямоугольной, а площадь (средний коэффициент силуэта) ближе к 1, тем лучше кластеризация. Внутри силуэта каждого кластера объекты расположены в порядке убывания их коэффициента силуэта, поэтому легко увидеть, какие именно объекты лучше соответствуют кластеру, а какие хуже. Значение silhouette index изменяется от -1 до 1:

Близкое к 1 — объект является похожим на другие объекты в кластере и не похожим на объекты из других кластеров.

Близкое к -1 — объект более подходит к соседним кластерам, чем к тому, в который он был распределён при кластеризации.

Близкое к 0 — объект расположен вблизи границы кластеров и высокой уверенности в его принадлежности нет.

Если большинство объектов имеют значения коэффициентов близкими к 1, можно утверждать, что кластерная структура хорошо выражена, и количество кластеров соответствует естественной группировке данных. Если в силуэте кластера много низких и отрицательных значений, это говорит о том, что кластерная структура плохо соответствует естественным группам данных, то есть кластеров слишком много или слишком мало.

Davies-Bouldin index показывает среднее «сходство» кластеров: расстояние между ними сравнивается с их размером. Чем меньше значение, тем лучше произведено разделение на кластеры. Davies-Bouldin index рассчитывается как отношение между расстояниями внутри кластера и между кластерами. Для каждого кластера вычисляется среднее расстояние между каждой точкой кластера и его центроидом (внутрикластерное расстояние), а затем сравнивается с расстоянием между центроидом этого кластера и его ближайшим кластером (межкластерное расстояние).

V-мера представляет собой гармоническое среднее значение homogeneity\_score и completeness\_score (между однородностью и полнотой). Метрика не зависит от абсолютных значений меток: перестановка значений меток классов или кластеров не влияет на значение индекса. Кроме того, она симметрична: замена true\_labels на predicted\_labels (действительных значений на прогнозируемые моделью) приведёт к тому же значению индекса. Это может быть полезно, чтобы измерить согласие двух независимых стратегий назначения меток на одном и том же наборе данных, когда реальная истинная метка неизвестна. Чем выше значение V-measure, тем выше качество кластеров.

Оценка результативности кластерного анализа исследуемого набора данных игровой активности во время соревнований по популярной киберспортивной игре League of Legends в зависимости от количества кластеров по описываемым выше метрикам представлена в таблице 2.

Таблица 2. – Оценка результативности кластерного анализа исследуемого набора данных в зависимости от количества кластеров по различным метрикам

| **Кол-во кластеров** | **Метрики оценки** | | |
| --- | --- | --- | --- |
| **silhouette** | **davies\_bouldin** | **vmeasure\_score** |
| 2 | 0.253960 | 1.655842 | 0.029570 |
| 3 | 0.227522 | 1.546734 | 0.053183 |
| 4 | 0.216540 | 1.534465 | 0.038606 |
| 5 | 0.224478 | 1.712410 | 0.306542 |
| 6 | 0.240833 | 1.576738 | 0.041028 |
| 7 | 0.237124 | 1.594108 | 0.210131 |
| 8 | 0.242359 | 1.479909 | 0.211322 |
| 9 | 0.253778 | 1.400488 | 0.156946 |
| 10 | 0.262073 | 1.413954 | 0.155082 |
| 11 | 0.266167 | 1.360035 | 0.212242 |

В результате можно сформулировать следующие основные цели кластерного анализа:

**Понимание данных, формулирование новых не вполне очевидных гипотез**. Деление разрозненных данных на группы помогает исследователю понять их сущность и выделить те подмножества, которые имеют в своей основе общий признак. Это позволяет подойти к каждому из таких подмножеств индивидуально и вместо общей закономерности на всё множество сформулировать набор правил для подмножеств. Также выделенные кластеры проще обрабатывать, применяя к разным кластерам различные методы анализа.

**Выявление аномалий в данных**. После кластерного анализа могут появиться отдельные объекты, которые сложно однозначно отнести к одному из определённых кластеров. Изучение подобных объектов может выявить как ошибки в измерениях или в самом эксперименте или позволит определить эти объекты как некий феномен, требующий отдельного изучения.

**Заполнение пропусков в данных**. Иногда при сборе данных у каких-то объектов будет зафиксирован не полный набор параметров и возникнут пустые ячейки в таблице. Разделение на кластеры поможет с большей точностью предположить отсутствующие признаки у других элементов кластера чем традиционное использование медианы или моды всего множества.

**Снижение размерности**. Если количество данных велико, то разделение их на кластеры позволит обрабатывать каждый кластер отдельно. Это позволит в дальнейшем использовать меньшие вычислительные мощности и удешевит анализ данных.

**Вопросы для контроля.**

Сущность алгоритма кластерного анализа k-means

Достоинства и недостатки алгоритма k-means

Выбор оптимального количества кластеров

Сущность алгоритма кластерного анализа DBSCAN

Достоинства и недостатки алгоритма DBSCAN

Метрики оценки результативности проведённого кластерного анализа

Основные цели кластерного анализа

# Список литературы

1. Aldenderfer M. S., Blashfield R. K. Cluster analysis sage university papers series //Quantitative applications in the social sciences. – 1984. – Т. 2.
2. Cattell R. B. The description of personality: Basic traits resolved into clusters //The journal of abnormal and social psychology. – 1943. – Т. 38. – №. 4. – С. 476-506
3. Driver H. E., Kroeber A. L. Quantitative expression of cultural relationships. – Berkeley: University of California Press, 1932. – Т. 31. – №. 4.
4. Duda R. O. et al. Pattern classification and scene analysis. – New York : Wiley, 1973. – Т. 3. – С. 731-739.
5. Estivill-Castro V. Why so many clustering algorithms: a position paper //ACM SIGKDD explorations newsletter. – 2002. – Т. 4. – №. 1. – С. 65-75.
6. Hall L. O., Ozyurt I. B., Bezdek J. C. Clustering with a genetically optimized approach //IEEE Transactions on evolutionary Computation. – 1999. – Т. 3. – №. 2. – С. 103-112.
7. Hand D. J. Principles of data mining //Springer-Verlag London Ltd. – 2007. –С. 530.
8. Hann J., Kamber M. Data mining: concepts and techniques. Morgan Kaufmann Publishers is an imprint of Elsevier – 2012. –740 р.
9. MacQueen J. Some methods for classification and analysis of multivariate observations //Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press. – 1967. - pp. 281–297.
10. Razmjooy N., Khalilpour M., Ramezani M. A new meta-heuristic optimization algorithm inspired by FIFA world cup competitions: theory and its application in PID designing for AVR system //Journal of Control, Automation and Electrical Systems. – 2016. – Т. 27. – С. 419-440.
11. Shirkhorshidi A. S. et al. Big data clustering: a review //Computational Science and Its Applications–ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30–July 3, 2014, Proceedings, Part V 14. – Springer International Publishing, 2014. – С. 707-720.
12. Zubin J. A technique for measuring like-mindedness //The Journal of Abnormal and Social Psychology. – 1938. – Т. 33. – №. 4. – С. 508-516
13. Ермаков, А. В. Моделирование темпо-ритмовой структуры соревновательной деятельности в рукопашном бою с использованием систем интеллектуального анализа данных / А. В. Ермаков // Экстремальная деятельность человека. – 2021. – № 3(61). – С. 38-42. – EDN UZJBED.
14. Ермаков, А. В. Прогнозирование с использованием методов математического моделирования в спорте высших достижений на примере зимних видов спорта / А. В. Ермаков, П. Е. Мякинченко // Теория и практика физической культуры. – 2021. – № 2. – С. 52-54. – EDN EXFWXQ.
15. Ермаков, А. В. Сравнительный анализ соревновательной деятельности в армейском рукопашном бое с использованием системы интеллектуального анализа данных ORANGE 3 / А. В. Ермаков, И. О. Яровой // Компьютерный спорт: проблемы и перспективы развития : Материалы Всероссийской научно-практической конференции, Москва, 02 декабря 2021 года. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский государственный университет физической культуры, спорта, молодёжи и туризма (ГЦОЛИФК)", 2021. – С. 45-51. – EDN SLTDSD.
16. Ермаков, А. В. Физкультурно-спортивное образование - педагогика на основе больших данных / А. В. Ермаков, Е. Н. Скаржинская, Е. А. Сарафанова // Большие данные в образовании: анализ данных как основание принятия управленческих решений : Сборник научных статей I Международной конференции, Москва, 15 октября 2020 года. – Москва: Издательский дом "Дело" РАНХиГС, 2020. – С. 313-323. – EDN WBZPIL.
17. Ермаков, А. В. Цифровая трансформация физкультурно-спортивного образования (на примере профессий «аналитик данных в ФКИС» и «тренер- аналитик») / А. В. Ермаков // Цифровая трансформация отрасли "физическая культура и спорт": теория, практика, подготовка кадров : Материалы Межрегионального круглого стола, Москва, 22 апреля 2021 года / Под редакцией М.А. Новоселова. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский государственный университет физической культуры, спорта, молодёжи и туризма (ГЦОЛИФК)", 2021. – С. 21-26. – EDN YFUMAM.
18. Ермаков, А. В. Цифровые технологии в спорте как инструмент воспитания / А. В. Ермаков, Е. Н. Скаржинская // Физическая культура и спорт в XXI веке: актуальные проблемы и их решения : Сборник материалов Всероссийской с международным участием научно-практической конференции, Волгоград, 21–22 октября 2020 года. Том 2. – Волгоград: Волгоградская государственная академия физической культуры, 2020. – С. 189-193. – EDN HUVOVT.
19. Ермаков, А. В. Использование кластерного анализа для определения стабильности состава технико-тактических средств грэпплинга на ADCC Submission Fighting World Championship / А. В. Ермаков // Боевые искусства и спортивные единоборства: наука, практика, воспитание : Материалы VIII Всероссийской научно-практической конференции с международным участием, Москва, 19 октября 2023 года. – Москва: Лика, 2023. – С. 100-105. – EDN WOJOUO.

1. Ермаков, А. В. Цифровые технологии в спорте как инструмент воспитания / А. В. Ермаков, Е. Н. Скаржинская // Физическая культура и спорт в XXI веке: актуальные проблемы и их решения : Сборник материалов Всероссийской с международным участием научно-практической конференции, Волгоград, 21–22 октября 2020 года. Том 2. – Волгоград: Волгоградская государственная академия физической культуры, 2020. – С. 189-193. – EDN HUVOVT. [↑](#footnote-ref-1)
2. Ермаков, А. В. Сравнительный анализ соревновательной деятельности в армейском рукопашном бое с использованием системы интеллектуального анализа данных ORANGE 3 / А. В. Ермаков, И. О. Яровой // Компьютерный спорт: проблемы и перспективы развития : Материалы Всероссийской научно-практической конференции, Москва, 02 декабря 2021 года. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский государственный университет физической культуры, спорта, молодёжи и туризма (ГЦОЛИФК)", 2021. – С. 45-51. – EDN SLTDSD. [↑](#footnote-ref-2)
3. Ермаков, А. В. Цифровая трансформация физкультурно-спортивного образования (на примере профессий «аналитик данных в ФКИС» и «тренер- аналитик») / А. В. Ермаков // Цифровая трансформация отрасли "физическая культура и спорт": теория, практика, подготовка кадров : Материалы Межрегионального круглого стола, Москва, 22 апреля 2021 года / Под редакцией М.А. Новоселова. – Москва: Федеральное государственное бюджетное образовательное учреждение высшего образования "Российский государственный университет физической культуры, спорта, молодёжи и туризма (ГЦОЛИФК)", 2021. – С. 21-26. – EDN YFUMAM. [↑](#footnote-ref-3)
4. Ермаков, А. В. Физкультурно-спортивное образование - педагогика на основе больших данных / А. В. Ермаков, Е. Н. Скаржинская, Е. А. Сарафанова // Большие данные в образовании: анализ данных как основание принятия управленческих решений : Сборник научных статей I Международной конференции, Москва, 15 октября 2020 года. – Москва: Издательский дом "Дело" РАНХиГС, 2020. – С. 313-323. – EDN WBZPIL. [↑](#footnote-ref-4)
5. Ермаков, А. В. Моделирование темпо-ритмовой структуры соревновательной деятельности в рукопашном бою с использованием систем интеллектуального анализа данных / А. В. Ермаков // Экстремальная деятельность человека. – 2021. – № 3(61). – С. 38-42. – EDN UZJBED. [↑](#footnote-ref-5)
6. Ермаков, А. В. Прогнозирование с использованием методов математического моделирования в спорте высших достижений на примере зимних видов спорта / А. В. Ермаков, П. Е. Мякинченко // Теория и практика физической культуры. – 2021. – № 2. – С. 52-54. – EDN EXFWXQ. [↑](#footnote-ref-6)
7. Razmjooy N., Khalilpour M., Ramezani M. A new meta-heuristic optimization algorithm inspired by FIFA world cup competitions: theory and its application in PID designing for AVR system //Journal of Control, Automation and Electrical Systems. – 2016. – Т. 27. – С. 419-440. [↑](#footnote-ref-7)
8. Ермаков, А. В. Использование кластерного анализа для определения стабильности состава технико-тактических средств грэпплинга на ADCC Submission Fighting World Championship / А. В. Ермаков // Боевые искусства и спортивные единоборства: наука, практика, воспитание : Материалы VIII Всероссийской научно-практической конференции с международным участием, Москва, 19 октября 2023 года. – Москва: Лика, 2023. – С. 100-105. – EDN WOJOUO. [↑](#footnote-ref-8)
9. Driver H. E., Kroeber A. L. Quantitative expression of cultural relationships. – Berkeley: University of California Press, 1932. – Т. 31. – №. 4. [↑](#footnote-ref-9)
10. Zubin J. A technique for measuring like-mindedness //The Journal of Abnormal and Social Psychology. – 1938. – Т. 33. – №. 4. – С. 508-516 [↑](#footnote-ref-10)
11. Cattell R. B. The description of personality: Basic traits resolved into clusters //The journal of abnormal and social psychology. – 1943. – Т. 38. – №. 4. – С. 476-506 [↑](#footnote-ref-11)
12. Estivill-Castro V. Why so many clustering algorithms: a position paper //ACM SIGKDD explorations newsletter. – 2002. – Т. 4. – №. 1. – С. 65-75. [↑](#footnote-ref-12)
13. Hall L. O., Ozyurt I. B., Bezdek J. C. Clustering with a genetically optimized approach //IEEE Transactions on evolutionary Computation. – 1999. – Т. 3. – №. 2. – С. 103-112. [↑](#footnote-ref-13)
14. Aldenderfer M. S., Blashfield R. K. Cluster analysis sage university papers series //Quantitative applications in the social sciences. – 1984. – Т. 2. [↑](#footnote-ref-14)
15. Duda R. O. et al. Pattern classification and scene analysis. – New York : Wiley, 1973. – Т. 3. – С. 731-739. [↑](#footnote-ref-15)
16. Hann J., Kamber M. Data mining: concepts and techniques. Morgan Kaufmann Publishers is an imprint of Elsevier – 2012. –740 р. [↑](#footnote-ref-16)
17. MacQueen J. Some methods for classification and analysis of multivariate observations //Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press. – 1967. - pp. 281–297. [↑](#footnote-ref-17)
18. Hand D. J. Principles of data mining //Springer-Verlag London Ltd. – 2007. –С. 530. [↑](#footnote-ref-18)
19. Shirkhorshidi A. S. et al. Big data clustering: a review //Computational Science and Its Applications–ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30–July 3, 2014, Proceedings, Part V 14. – Springer International Publishing, 2014. – С. 707-720. [↑](#footnote-ref-19)