МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ И РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение

высшего образования

**«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

А.А. РОМАНОВ

**Анализ многомерных данных**

**Учебно-методическое пособие**

Ульяновск

УлГТУ

2023

**УДК 004.9 : 338(076)**

**ББК 32.973.26 – 018.2я73**

|  |  |
| --- | --- |
| К69  К69 | Рекомендовано научно-методической комиссией факультета информационных систем и технологий в качестве практикума.  **Романов, Антон Алексеевич**   АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ : лабораторный практикум / А. А. Романов. – Ульяновск : УлГТУ, 2023. – 36  с.  Учебно-методическое пособие адресовано студентам для изучения дисциплины «Анализ многомерных данных». Рекомендации и требования к лабораторным работам разработаны в соответствии с рабочей программой дисциплины. Пособие предназначено для студентов, обучающихся по направлениям связанным с разработкой и использованием методов искусственного интеллекта.  Работа подготовлена на кафедре «Информационные системы». |

**УДК 004.9 : 338(076)**

**ББК 32.973.26 – 018.2я73**

© Романов А.А., 2023

© Оформление. УлГТУ, 2023

СОДЕРЖАНИЕ

[Введение 4](#_Toc150355105)

[Лабораторная работа №1 13](#_Toc150355106)

[Лабораторная работа №2 21](#_Toc150355107)

[Лабораторная работа №3 25](#_Toc150355108)

[Лабораторная работа №4 31](#_Toc150355109)

[Лабораторная работа №5 35](#_Toc150355110)

[Лабораторная работа №6 39](#_Toc150355111)

[Список литературы 47](#_Toc150355112)

[Приложение 1. Временные ряды NN3 48](#_Toc150355113)

## Введение

Целью освоения дисциплины «Анализ многомерных данных» является освоение методов, исследующих зависимости в данных, описываемых несколькими переменными.

Задачами освоения дисциплины (модуля) являются:

* формирование навыков использования классификации методов многомерного анализа данных;
* формирование навыков анализа факторов;
* формирование навыков проектирования программных систем и программных комплексов, решающих задачи построения прогнозов;
* формирование навыков создания программных проектов, реализующих интеллектуальные методы анализа многомерных данных;
* изучение практических аспектов внедрения реализованных методов анализа многомерных данных в различных предметных областях;
* формирование навыков экспериментального обоснования выбранного подхода анализа данных.

Дисциплина «Анализ многомерных данных» рассматривает методы и модели, которые требуют для предиктивной аналитики. Для того, чтобы правильно позиционировать эту дисциплину, необходимо обозначить ограничения рассматриваемых концепций.

Во-первых, если говорить вообще про сферу искусственного интеллекта, то существуют и используются различные типы данных, достаточно разнообразных, имеющих свою специфику и отличающихся от традиционных для области программных систем типов данных. В первую очередь стоит вспомнить методы для анализа больших массивов данных, область исследования которых лежит в области анализа больших данных.

Во-вторых, есть специфичные сведения, которые попадают в категорию неструктурированных данных. Это все то, что касается различного представления текстовых массивов, массивов сведений, которые содержатся в видео и различных других хранилищах.

Такие данные являются продуктом работы множества информационных систем и информационных процессов. Следовательно, нужно учитывать специфику выбора методов для анализа данных. В первую очередь стоит обратить внимание на структуру самих данных, историю их накопления. Часто перед началом работ требуется выполнить их нормализацию, предобработку, придать им структурированный вид в зависимости от решаемой задачи. К таким структурированным типам данных, которые могут существовать в различных проблемных областях, относятся, например, временные ряды. Эта структура позволяет решать широкий круг задач, отражающих динамику изменения показателей некоторого объекта или объектов проблемной области. Исследования в области временных рядах породили большое число методов, в том числе интеллектуальных, гибридных.

Если затронуть другие типы данных, например массивы изображений, видеопотоки, то можно отметить, что с ними хорошо работают различные методы машинного обучения: для выявления объектов, распознавания сцен, последующей интерпретации, оценки состояния объектов.

Извлечение полезных закономерностей при помощи вышеописанных структур данных развивает область data mining, и, в частности, time-series data mining, то есть это извлечение полезных сведений из массивов временных рядов. Временные ряды порождаются в системах, в которых есть некоторая историчность, то есть имеется совокупность показателей, значения которых можно рассматривать с аргументом времени.

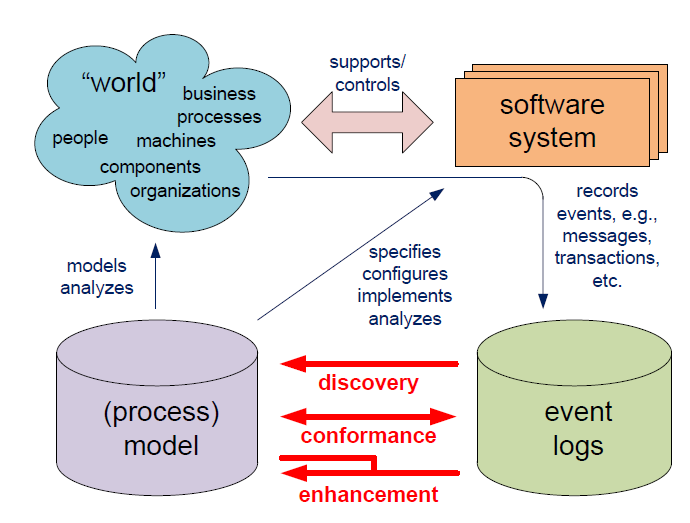
Что может служить источником таких сведений? В первую очередь необходимо определить, для каких целей производится анализ. Часто необходимо на основе изучения истории изменений в какой-либо системе понять, как лучше ей управлять. Для этого необходимо построить процессную модель, отражающую переходы системы между состояниями. Этот процесс называется process-mining - выявление и анализ процессов в ходе исследования каких-либо журналов, событий, журналов, информационных и в любых других программных систем.

С точки зрения дисциплины «Анализ многомерных данных» подразумевается, что имеется собственная специфика данных и что можно применить подходы data-minig для зрения решения задач.

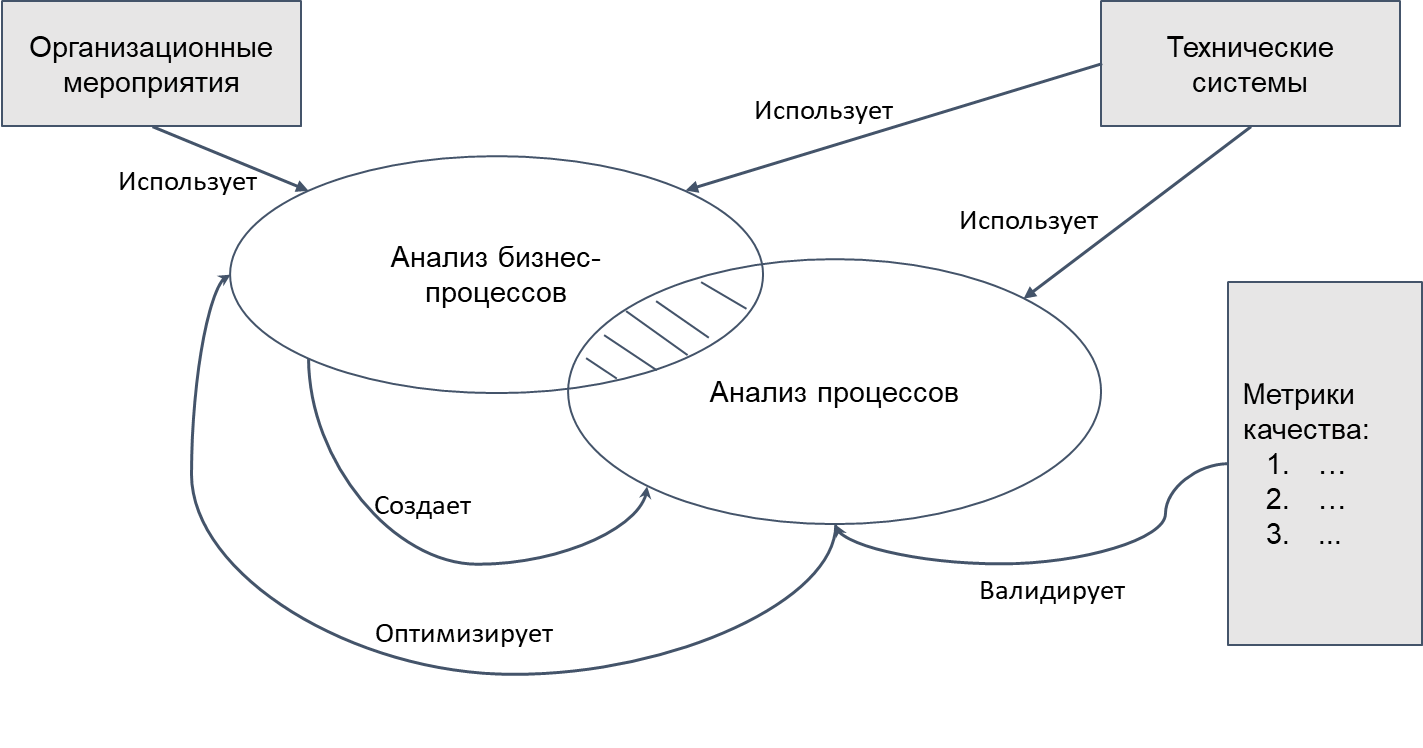
Интеллектуальный анализ процессов - это мост между интеллектуальным анализом данных и моделированием / анализом процессов. Интеллектуальный анализ процессов и интеллектуальный анализ данных имеют много общего. Большинство методов интеллектуального анализа данных, таких как классификация, кластеризация и анализ последовательностей, также могут использоваться в интеллектуальном анализе процессов.

Методы интеллектуального анализа процессов включают три основных класса:

* обнаружение процессов (автоматическое обнаружение модели процесса из журнала событий),
* проверка соответствия (обнаружение отклонений путем сравнения модели процесса и соответствующего журнала событий),
* организационного анализа (включая несколько методов, таких как как системы анализа, прогнозирования и рекомендаций социальных сетей).



В то же время выявленные анализы процессов позволяют, например, оптимизировать бизнес-процесс, то есть это выдать конкретные рекомендации проведения или переорганизации мер-приятий, коммуникации, возможно, выявить какие-то узкие места при организации какого-либо процесса обработки данных или просто использовании самой системы. И соответственно, с точки зрения использования, как правило, такие рекомендации позволяют, например, перестроить технические системы, позволяют настроить сценарии работы в них дополнительные.



Методы интеллектуального анализа процессов часто работают с журналами событий в качестве входных данных. Каждые данные события, которые можно использовать, должны иметь как минимум три свойства:

* данные должны иметь временные метки;
* должны присутствовать метки действий;
* должен быть указан идентификатор случая для каждой записи (идентификатор случая - это идентификатор каждого экземпляра процесса).

Меры качества в анализе данных

* Пригодность: этот критерий показывает, насколько обнаруженная модель процесса соответствует журналу событий. Распространенный способ вычисления пригодности - это воспроизведение всех трассировок в модели процесса и подсчет количества случаев, когда трассировка не может быть воспроизведена в модели процесса. Модель процесса, которая слишком приспособлена, может страдать от проблемы переобучения.
* Простота: чем более проста открытая модель, тем она востребованнее. Различные способы вычисления простоты основаны на количестве действий и отношений.
* Точность: модель является точной, если не допускает слишком большого количества ненаблюдаемого поведения в журналах событий.
* Обобщение: этот критерий используется, чтобы избежать проблемы переобучения. Модель должна иметь минимальное обобщение и не ограничиваться поведением, наблюдаемым в журналах, т.к. должно учитываться, что журналы событий могут быть неполными.

Первый основной тип и одна из самых сложных задач интеллектуального анализа процессов - это обнаружение процессов. Метод обнаружения процессов принимает журнал событий информационной системы в качестве входных данных и генерирует модель без использования какой-либо априорной информации. Обнаружение процесса можно исследовать с различных точек зрения, например, с точки зрения потока управления, с точки зрения организации работ, с точки зрения сценария и с точки зрения временных характеристик.

В качестве результата работы алгоритмов извлечения можно получить модель процесса в одном из формализмов: сеть Петри, сеть потоков работ (WF-сеть), сеть BPMN, эвристическая сеть и другие.

Второй важный тип анализа процессов - это проверка соответствия. Проверка соответствия используется для систем обнаружения отклонений, прогнозирования, принятия решений и рекомендаций. При проверке соответствия журнал событий сравнивается с существующей соответствующей моделью процесса, и он показывает, что если модель процесса соответствует реальности, и наоборот.

Для проверки соответствия журнала событий применяется модель, которая последовательно выполняет алгоритм. Изначально устанавливается начальная разметка модели с одним маркером в начальной позиции. Далее, алгоритм проходит по каждой трассе журнала и для каждого события предусмотрено два варианта: возможен переход с соответствующим именем или нет. В случае срабатывания перехода, разметка модели меняется и алгоритм переходит к следующему событию журнала. Если же переход не активен, то в некоторых позициях необходимых для его выполнения, может не быть достаточного количества маркеров. В таком случае, в рассматриваемую позицию добавляется искусственный маркер, который позволяет модели продолжить исполнение. Такая позиция помечается как содержащая недостающие маркеры и при возникновении подобной ситуации количество недостающих маркеров увеличивается.

Сбор, анализ и подготовка многомерных данных.

Необходимо отметить актуальность задачи сбора и анализа многомерных данных. Не все сведения, которые мы собираем и извлекаем из сторонних источников, могут быть одновременно нужны при проведении анализа. Особенно это может касаться источников больших данных. Характеристики исследуемых объектов не всегда могут быть представлены в количественном виде, часто бывает, что они оцениваются качественно.

Выборка может представлять совокупность сведений по целому ряду объектов. И эти объекты никаким образом не описаны, а именно мы можем затрудняться в определении того, какой из атрибутов какому объекту относится.

* Многие прикладные задачи в настоящее время сводятся к анализу многомерных множеств.
* В общем виде они представляются как набор количественных характеристик объектов.
* Изучая наборы характеристик можно сделать предположение о разделении объектов на группы (категории, классы).
* Результатами исследования таких категорий является заключение о схожих закономерностях поведения объектов, в том числе и взаимном влиянии групп друг на друга.

Многомерное пространство характеристик - совокупность свойств, одновременно приписываемых изучаемому объекту. Изучение объекта происходит через совокупность изучаемых (выделенных) количественных характеристик. Алгоритм можно описать следующей последовательностью шагов:

* Сбор данных о характеристиках.
* Отбор характеристик. Сочетание характеристик может искажать представления о закономерностях. Поэтому следует выбирать наиболее значимые.
* Построение многомерного пространства. Процедура включает формирование шкал, их масштабирование и нормирование значений характеристик. Масштабирование сильно влияет на значение мер расстояний между объектами в пространстве.
* Проведение структурного или визуального анализа. Здесь важно решить задачу визуализации n-мерного пространства. Отдельно может быть решена задача сокращения размерности пространства.

В системах обработки информации данные подразделяются на:

* качественные (нечисловые);
* количественные (числовые).

Предварительно перед анализом их следует преобразовать с использованием шкал.

**Номинальная шкала** используется для разбиения измерений на классы в соответствии с определенным признаком, где каждому классу присваивается одно обозначение. Хотя арифметические операции не имеют смысла для таких классов, возможны количественный анализ, например, подсчет числа объектов и их частоты встречаемости.

В **порядковой шкале** классы не только разбиваются на качественные измерения, но и упорядочиваются в соответствии с важным порядком следования классов для сопоставления элементов внутри них.

**Количественные измерения** могут быть интервальными, относительными, разностными и абсолютными значениями. Для всех шкал количественных измерений допустимы арифметические операции.

На практике объекты могут иметь множество различных свойств и связей между ними, как например, объекты в социологии, эконометрике, медицине, экологии и сложных технических системах.

На практике объекты характеризуются большим количеством различных свойств и связей между ними, примерами являются объекты в социологии, эконометрике, медицине, экологии, сложных технических системах.

Многомерные данные можно представить как матрицу, связывающую объекты и их признаки: в строках представлены анализируемые объекты, а в столбцах – значения изучаемых признаков.

Особенностью и ограничением такого представления является то, что значения признаков попадают в матрицу с учетом времени. Т.е. для объектов с большим числом характеристик нужно помнить, что возможно есть временной сдвиг между собранными значениями и выявлять закономерности в такой матрице нужно с таким учетом. Либо делается предположение, что за отведенный интервал времени сам объект не изменяет своего состояния.

В многомерном анализе, временные ряды являются одним из важных типов данных. Они представляют собой последовательность значений, упорядоченных во времени, и отражающие уровень состояния и изменения показателя. Однако, каждое значение в ряду зависит от предыдущих состояний, и порядок значений имеет важное значение. Временные ряды отличаются по длине и характеру изменения времени в процессе их наблюдения. Количество информации, также, не пропорционально размеру выборки и последовательные значения не независимы.

Характеристики временных рядов включают:

1. Длину ряда, которая может отражать количество наблюдений параметра или интервал времени между начальным и конечным моментом наблюдения. В отличие от обычного статистического анализа, информация во временных рядах не зависит от размера выборки, и последующие значения не являются независимыми.
2. Дискретность или непрерывность ряда определяется тем, как изменяется параметр во время наблюдения. Дискретные ряды могут быть получены из непрерывных путем выборки через интервалы или накопления значений за определенный период времени.
3. Определение детерминированности временных рядов зависит от того, можно ли представить будущие значения ряда в виде математической функции. Если вероятности используются для описания будущих значений, то ряд считается недетерминированным или случайным. Стохастические временные ряды могут быть стационарными или нестационарными, а стационарные ряды имеют постоянные свойства во времени, не зависящие от начальной точки.
4. Временные ряды могут быть интервальными, когда меряют изменение явления за определенный промежуток времени, или моментными, когда они отображают уровень явления в конкретный момент времени. Моментные ряды отличаются от интервальных в том, что сумма уровней интервального ряда дает объективную меру за интервал, в то время как моментный ряд характеризует состояние в конкретный момент времени.
5. Различают полные и неполные временные ряды в зависимости от того, следуют ли даты регистрации или окончания периодов друг за другом с равными интервалами.

Изучение временных рядов очень важно для выявления закономерностей и зависимостей их изменения на протяжении длительного периода. Однако это сложный процесс, так как данные зависимости формируются под воздействием множества факторов. Выделяется две группы:

* определяющие основную тенденцию динамики;
* вызывающие случайные колебания.

Временной ряд можно выразить как сумму двух величин:

Величина является случайной величиной, не автокоррелированной и со средним значением равным нулю, а может быть как детерминированной функцией, так и случайным процессом, либо их комбинацией. Разница между и заключается в том, что влияет только на значение текущего периода, тогда как может воздействовать на несколько или все последующие периоды. Основная тенденция или тренд определяет характер изменения явления в течение длительного времени, без учета случайных факторов. Тренд может быть детерминированным или случайным и обозначается через , а колебания — это отклонения уровней от тренда. Можно наблюдать иерархию тенденций и колебаний: та величина, которая для, например столетия, выступает как колебания, на интервале времени низшего порядка, например трех-пяти лет, может выступать как тенденция. Временной ряд можно представить в виде комбинации четырех компонент: тренда, регулярных колебаний вокруг тренда, сезонности и случайного эффекта. Для изучения временного ряда необходимо разложить его на составляющие эти компоненты, которые являются теоретическими величинами и не могут быть наблюдаемыми. Разделение временного ряда на эти компоненты является основной задачей анализа временных рядов.

Модель является главным инструментом для анализа и прогнозирования временных рядов. Ее удобно использовать в дискретных системах, где можно наблюдать и регулировать процессы через определенные интервалы времени. Термин "модель" имеет два значения: как описание процесса генерации временного ряда и как инструмент для прогнозирования. Основное различие между ними заключается в том, что модель временного ряда показывает фактические значения, а прогнозная модель дает оценку будущих значений временного ряда.

## Лабораторная работа №1

Тема лабораторной работы: подготовка и оценка набора многомерных данных.

**Методические рекомендации и материалы**

Для выбора набора данных можно обратиться к агрегаторам, которые предоставляют ссылки на открытые наборы данных. Примером такого агрегатора является сайт proglib: https://proglib.io/p/top-30-resursov-s-dannymi-dlya-mashinnogo-obucheniya-2022-03-31 , здесь в качестве источника для лабораторной работы был выбран источник из раздела «Датасеты общего назначения» - «Министерство Культуры РФ»: <https://opendata.mkrf.ru/opendata/7705851331-national_film_identities>. Набор данных представляет из себя «Реестр удостоверений национального фильма», представляет из себя 63 столбца и 36207 строк, в которых содержится информация о фильмах, соответствующих статусу национального фильма, их параметрах, продюсерах, авторах, сроках действия удостоверений. Примеры полей:

* Идентификатор записи реестра,
* Название фильма,
* Вид фильма,
* Количество серий,
* Формат,
* Цвет,
* Фонограмма,
* Объем фильма,
* Автор сценария,
* Режиссер,
* Композитор,
* Продюсер,
* Правообладатель,
* Юридический адрес продюсера (по месту регистрации),
* Продюсер является юридическим лицом,
* Начало производства фильма,
* Окончание производства,
* Дата выдачи удостоверения,
* Срок действия удостоверения,
* Стадия производства,
* Коллекция,
* Сметная стоимость,
* Тип инвестирования.

Выполним все преобразования в Google Colab. Первым шагом импортируем все библиотеки и методы, которые потребуются для выполнения работы в нашей среде (рисунок 1).

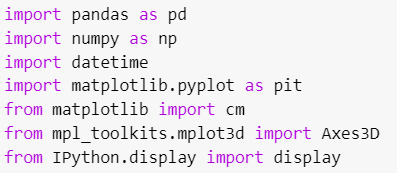


Рисунок 1. Импортируемые библиотеки

Добавим в проект несколько функций для визуального отображения результатов обработки (рисунок 2).



Рисунок 2 – функции для визуального отображения результатов

Теперь попробуем прочитать исходный набор данных и вывести количество выпущенных фильмов, сгруппировав их по полю «Вид фильма». Для этого запустим код на рисунке 3, и получим результат на рисунке 4.

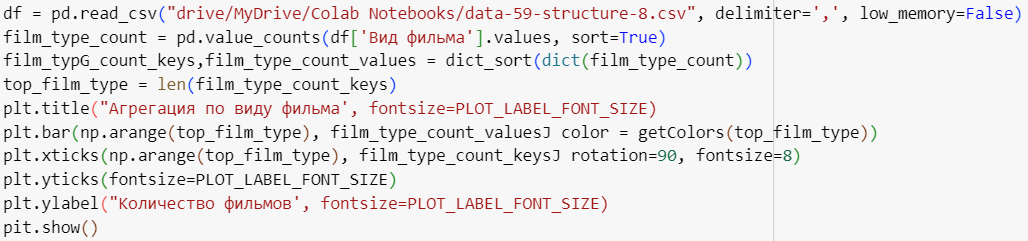


Рисунок 3 – код для агрегации

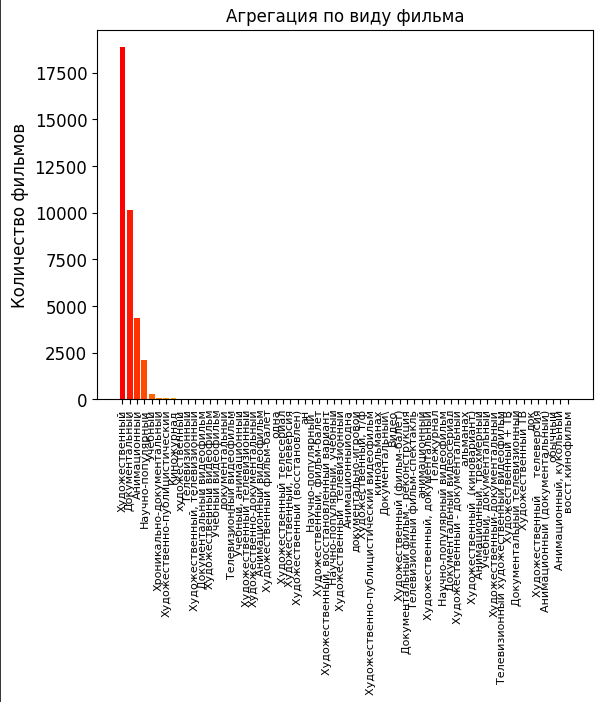
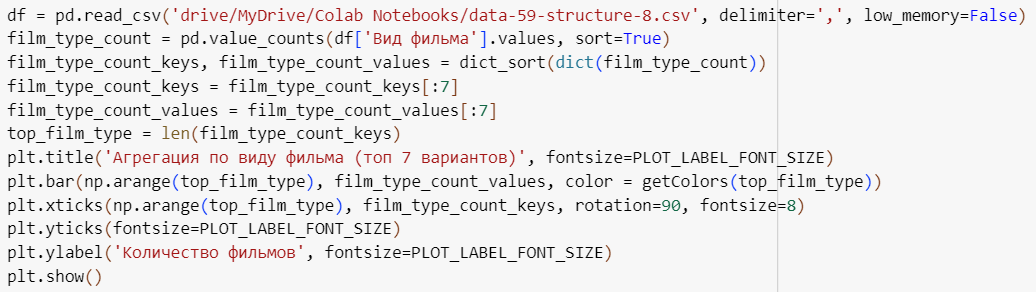


Рисунок 4 – результат агрегации

На рисунке мы видим, что в массиве данных присутствует большое количество уникальных значений в поле «Вид фильма», что сильно усложняет возможность проанализировать массив. Чтобы исправить эту ситуацию можно все виды фильмов, которые не имеют большого влияния на результат объединить в одну группу «Другой».

Чтобы понять какие виды фильмов оставить, выполним код на рисунке 5, который оставит только первые 7 самые крупные группы видов фильмов, и посмотрим на результаты на рисунке 6.

Рисунок 5 – код топ 7 групп фильмов

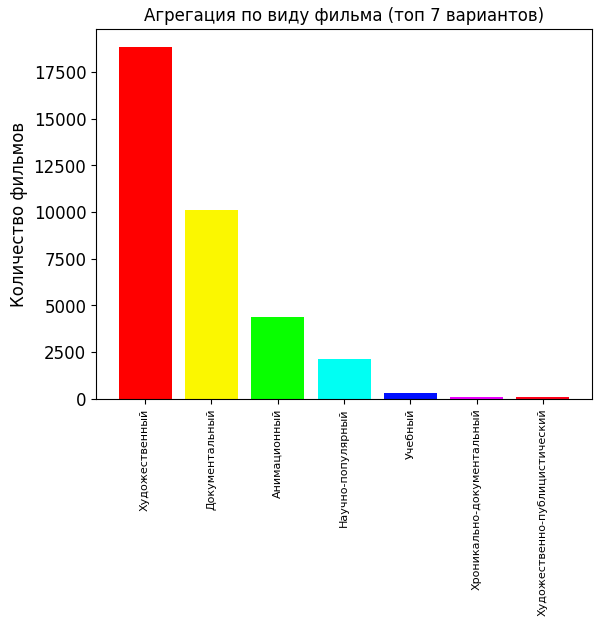


Рисунок 6 – топ 7 видов фильмов

Теперь объединим все виды после «Научно-популярный» в группу другие (рисунок 7,8).

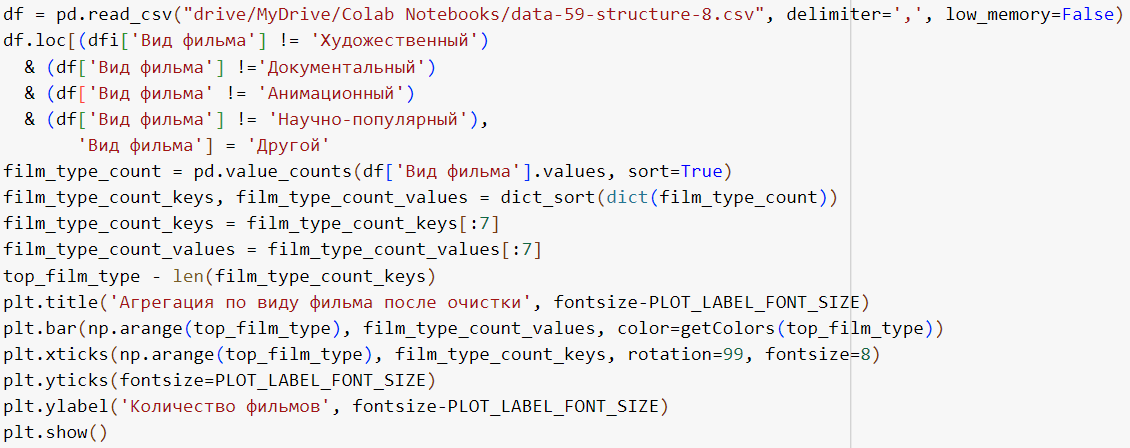


Рисунок 7 – код для группировки

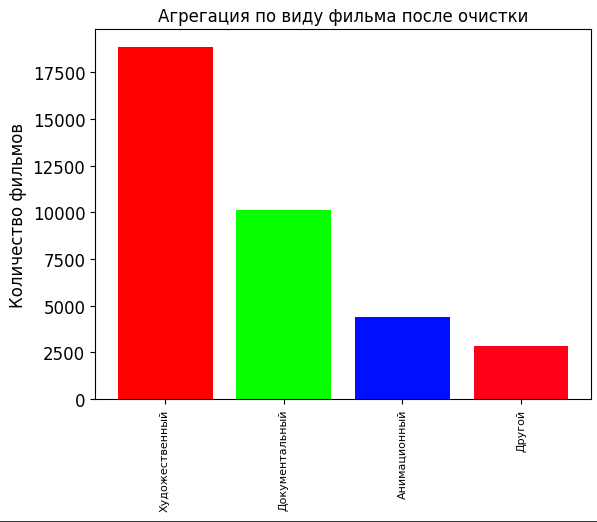
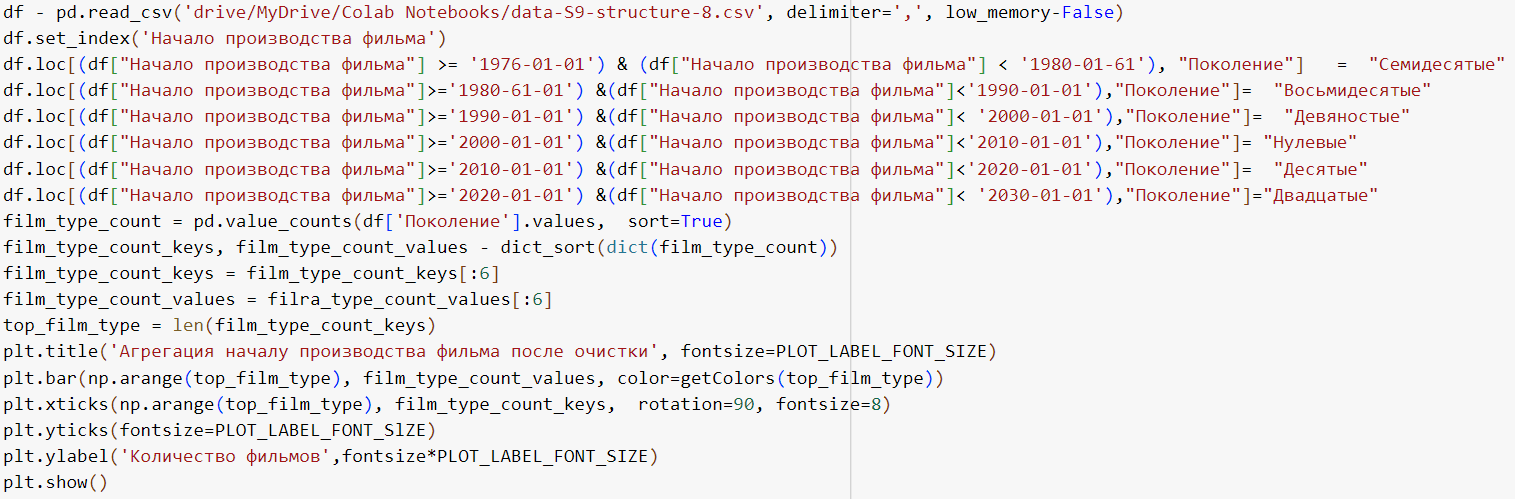


Рисунок 8 – результат группировки

После проведенной очистки данных, глядя на результаты, можно делать первые выводы (извлекать ценность) из многомерного массива данных.

Альтернативным важным параметром исходного массива может являться поле «Начало производства фильма». Попробуем объединить значения в группы по десятилетиям, и покажем результаты по количеству фильмов, которые начали снимать по десятилетиям (рисунок 9, 10).

Рисунок 9 – код для группировки

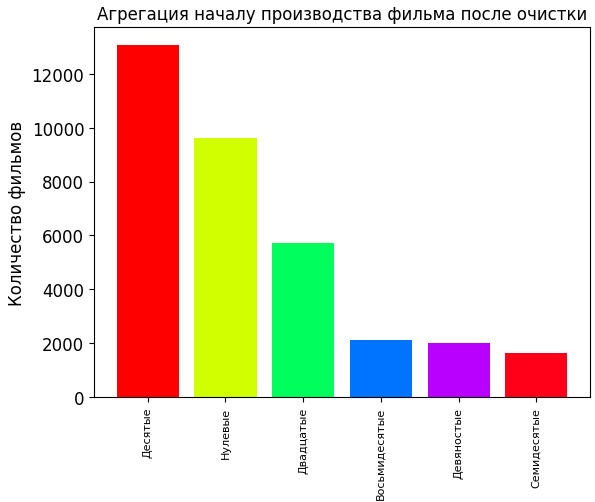


Рисунок 10 – результат обработки

Самым удачным качественным значением в исходном массиве данных является группировка по количеству фильмов. Рисунок 8 и 10 отображают самые лучше качественные выводы, которые можно сделать по итогам проведенного анализа.

**Задания к лабораторной работе**

1. Найти открытый набор многомерных данных.
2. Выделить из всего множества атрибутов самые значимые для дальнейшего анализа.
3. Нормализовать данные атрибутов объектов.
4. Выбрать шкалу для преобразования качественных значений и представить полученный набор данных
5. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание набора многомерных данных.
   3. Описание решения поставленных задач.
   4. Список использованных источников.
6. Оформить и защитить отчет.

**Контрольные вопросы**

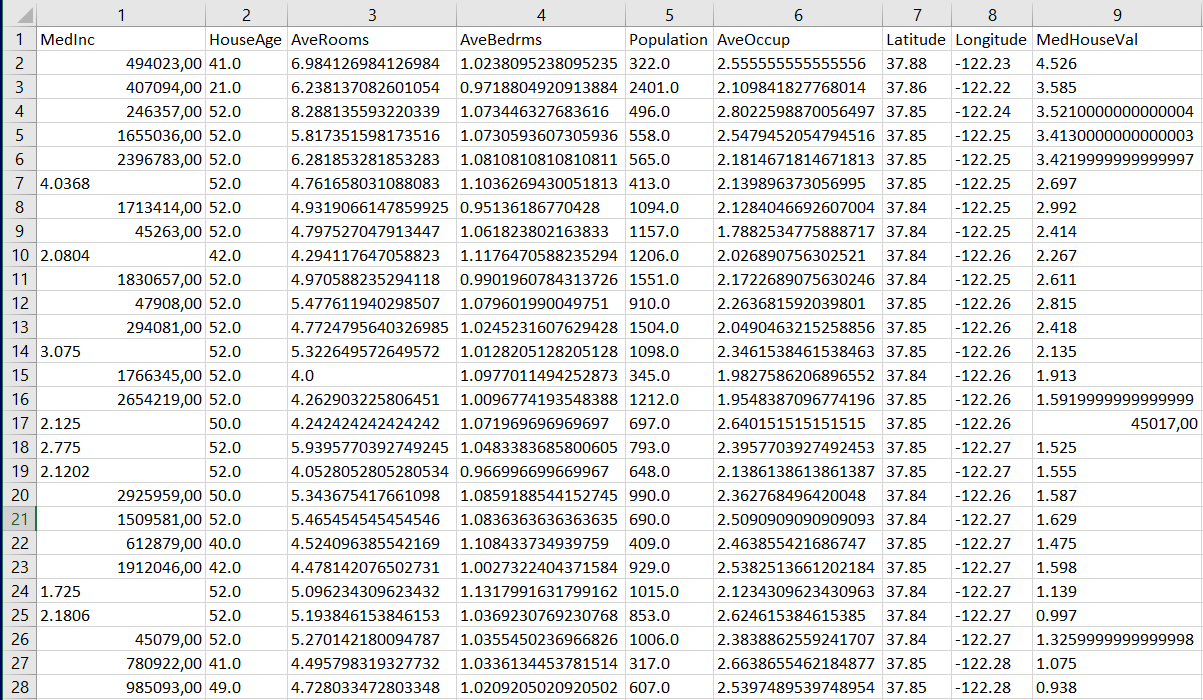
1. В чем состоит нормализация данных?
2. Каковы проблемы анализа многомерных данных?
3. Приведите примеры шкал для преобразования значений многомерных данных.
4. Какими принципами можно руководствоваться при выборе атрибутов многомерных данных для решения задач анализа?
5. Какова структура многомерных данных?

## Лабораторная работа №2

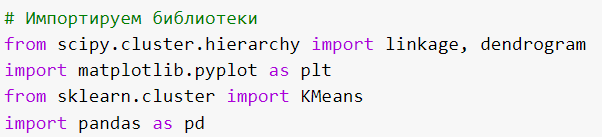
Тема лабораторной работы: кластеризация многомерных данных.

**Методические рекомендации и материалы**

Анализ данных может быть осложнен неподготовленными данными. Данные могут быть неструктурированными, храниться в разных источниках, иметь разные форматы. Для снижения размерности задачи анализа, выделения значимых групп анализируемых объектов применяю методы кластеризации. Эти методы, основываясь на признаках объектов, позволяют расположить объекты в n-мерном пространстве относительно друг друга. Методы кластеризации имеют реализации в библиотеках для python. Для их использования необходимо ознакомиться с документацией, подготовить данные в нужном виде, задать необходимые параметры и выполнить кластеризацию.

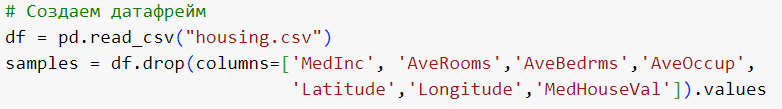


**Рисунок 1. Пример данных для кластеризации**

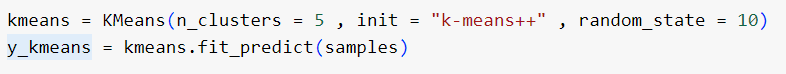


**Рисунок 2. Подключаемые библиотеки**

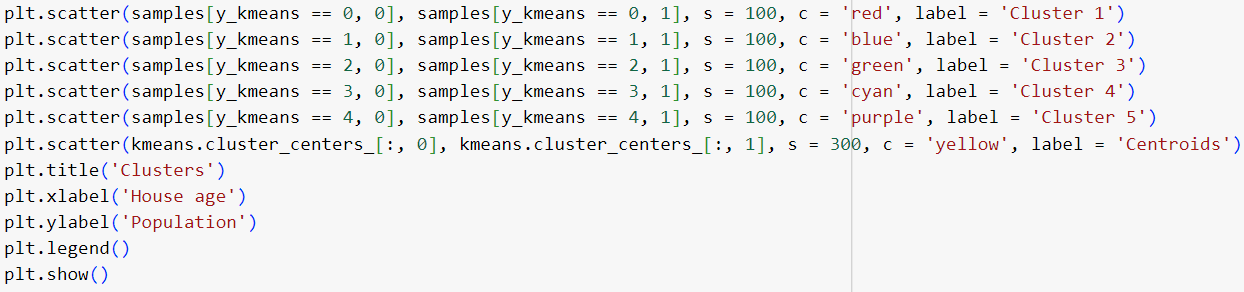
Данные для кластеризации могут быть загружены из структурированных файлов, например csv. В этом случае необходимо выполнить фильтрацию атрибутов, т.к. выгруженные сведения могут потребоваться не все, а только те, на основе которых будет производиться кластеризация. Пример такой предобработки приведен на рисунке 3.

**Рисунок 3. Загрузка данных**

Методов кластеризации существует большое количество. Применяются они в соответствии со своим назначением: получение иерархии множества связанных объектов, задание требуемого числа кластеров, нечеткость связи между объектами в кластере, применение различных мер расстояний в пространстве кластеризации и т.д. В данном примере рассмотрен метод KMeans. Его особенностью является то, что он позволяет кластеризовать все объекты на заранее заданное число кластеров. Пример приведен на рисунке 4.

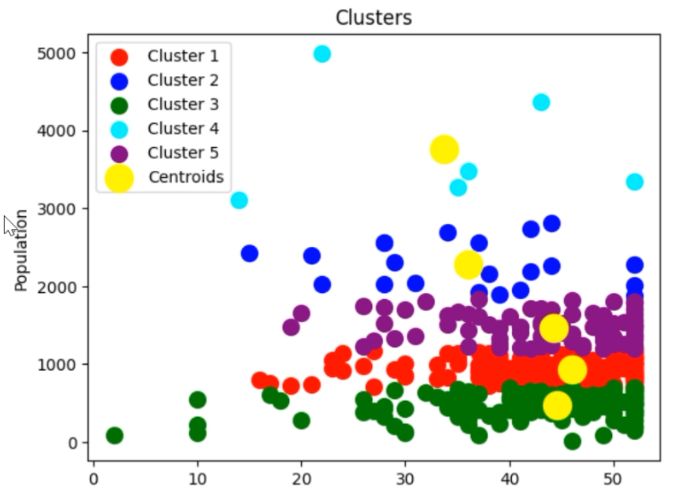
**Рисунок 4. Пример инициализации метода кластеризации**

Одними из трудоемких процедур при кластеризации являются извлечение и подготовка данных, описание кластеров и интерпретация результатов. Проблема визуализации многомерных больших даннных отмечается исследователями особенно остро, например в статье «A Clustering Visualization Method for Density Partitioning of Trajectory Big Data Based on Multi-Level Time Encoding» авторов Boan Wei, Jianqin Zhang, Chaonan Hu, Zheng Wen. Вариантов представления результатов кластеризации может быть несколько: двумерные диаграммы, трехмерные диаграммы, деревья объектов,



**Рисунок 5. Пример разметки кластеров**

Результаты кластеризации приведены на рисунке 6.



**Рисунок 6. Результаты кластеризации на основе двух признаков.**

**Задания к лабораторной работе**

1. Выбрать один из методов кластеризации.
2. Реализовать скрипт для кластеризации данных.
3. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание метода кластеризации.
   3. Описание решения поставленных задач.
   4. Описание полученных кластеров.
   5. Список использованных источников.
4. Оформить и защитить отчет.

**Контрольные вопросы**

1. Назовите методы кластеризации для многомерных данных.
2. Какие параметры методов могут быть использованы при кластеризации?
3. Какие параметры многомерных данных влияют на качество кластеризации?
4. Как можно измерить качество кластеризации?
5. Каковы основные этапы реализации процедуры кластеризации в python?

## Лабораторная работа №3

Тема лабораторной работы: Корреляционный анализ многомерных временных рядов

**Методические рекомендации и материалы**

Анализ многомерных временных рядов может быть выполнен для выявления закономерностей во всем множестве атрибутов, что является особенно полезным, потому что такой подход может предоставить больше информации, чем анализ отдельного временного ряда. В этой связи необходимо выделить какие операции могут быть совершены над множество временных рядов:

* Кластеризация временных рядов для использования центров кластеров для смыслового описания выделенных временных рядов – атрибутов многомерных данных.
* Группирование временных рядов для расширения сведений о закономерностях отдельных временных рядов.
* Фильтрация временных рядов для выявления причинно-следственной связи между рядами.

Если принимать во внимание приложение временных рядов к экономике, часто имеются показатели, изменения которых характеризуются большой степенью сходства. Это говорит и о похожем характере поведений локальных и глобальных тенденций временных рядов. На примере такой предметной области легко можно понять, что многие процессы могут изменяться синхронно. Проводя анализ с целью выявления групп похожих временных рядов показателей можно выявлять информацию об отношении каждого показателя к значимому фактору. Положительными эффектами от такого анализа являются:

* Сокращение времени прогнозирования и упрощение прогнозных операций в целом за счет упрощения вычислительных операций.
* Выявление групп похожих процессов в общем массиве с целью получения информации о факторах, группирующих эти процессы.

Определим понятия, которые будут встречаться в тексте. Похожие временные ряды — ряды, имеющие коэффициент корреляции в рамках заданного порога. Группа временных рядов – множество временных рядов, временные ряды которого имеют значение коэффициента совместной корреляции, лежащее в указанных пределах. В данном материале будет использован подход F-преобразования временных рядов.

Для каждого временного ряда задаются функции нечеткого разбиения, строится F-преобразование нулевого порядка (F0-преобразование). Для построения F-преобразования первого порядка [5] (F1-преобразование): , необходимо определить коэффициенты ,. В формуле k=1,…,n, n – количество компонент для ряда. Коэффициенты являются компонентами F0-преобразования. Коэффициенты определяются как:

,

где - значения временного ряда в момент времени , - базисные функции нечеткого разбиения для F1-преобразования в момент времени . Kоэффициенты равны средневзвешенному значению тангенса угла наклона кусочно-линейного тренда временного ряда. Именно они будут служить основой анализа групп временных рядов. Сами компоненты F1-преобразования представляют собой векторы и описывают локальные тенденции временного ряда. В качестве примера построения одной из таких компонент приводим график временного ряда №104 из NN3 (см. приложение 1) и компоненту последнего отрезка разбиения :

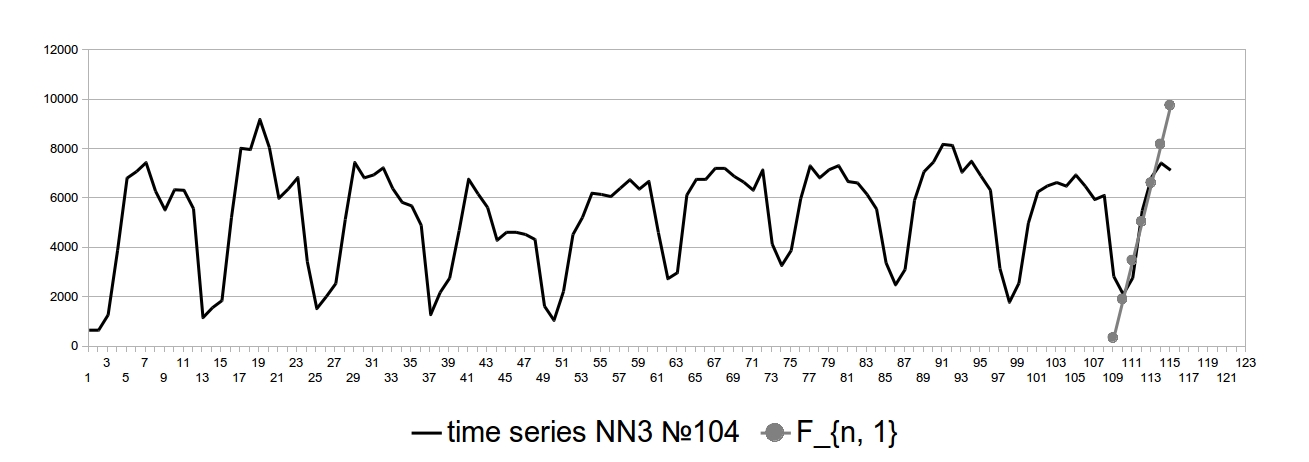


Рис. 1. ВР №104 из NN3 и компонента последнего отрезка разбиения .

Группирование временных рядов будем производить, анализируя коэффициенты . Для каждой пары временных рядов вычисляется коэффициент корреляции Пирсона (линейный коэффициент корреляции):

,

где значения коэффициентов для первого и второго временного ряда из пары соответственно.

Важным условием для проведения анализа является равная длина временных рядов. Перед построением компонент F-преобразования различного порядка производится нормировка значений временного ряда к интервалу [0,...,1] по модулю. Нормировка необходима для обеспечения одинаковой области значений коэффициентов . При формировании групп временных рядов возникают следующие проблемы.

Необходимо правильно выбрать количество точек, которая покрывает базисная функция — данная величина непосредственно влияет на последующую корреляцию временных рядов по коэффициенту . Чем большее количество точек будет покрывать базисная функция, тем более сглаженным будет ряд, и корреляция будет рассчитываться по общим тенденциям рядов.

Задание порога влияет на количество похожих по коэффициенту корреляции временных рядов. Важно определить, что более важно: включить в группу больше рядов, или отобрать только самые близкие по корреляции ряды.

Один из вариантов формирования группы по данным результатам - выделить ряды, имеющие максимально количество похожих и включить в список все участвующие ряды. Например, можно выделить следующую группу рядов: 1,2,3,7,8,9,10,11,12,13,14.

|  |  |
| --- | --- |
| 1 |  |
| 2 |  |
| 3 |  |
| 7 |  |
| 8 |  |
| 9 |  |
| 10 |  |

Рис. 2. Представители рядов, вошедшие в одну группу.

Не вошли в группу похожих рядов:

|  |  |
| --- | --- |
| 4 |  |
| 5 |  |
| 6 |  |

Сформированные группы временных рядов позволяют упростить построение прогноза за счет сокращения вычислительных операций и идеи использования спрогнозированных локальных тенденций одного ряда для построения прогноза других рядов группы.

Сложностями при группировании временных рядов являются:

* Наличие сдвига зависимостей во временных рядах. Предварительно необходимо выявлять временной отступ до начального момента взаимной синхронизации зависимости во временных рядах.
* Наличие псевдозависимостей. Алгоритм выявления коррелированных временных рядов может однозначно показать, что взаимосвязь присутствует. Однако на самом деле это может быть просто совпадением. В этом случае требуется дополнительный анализ.

**Задания к лабораторной работе**

1. Для одного из соревнований по прогнозированию выбрать временные ряды из разных категорий.
2. Выбрать метод корреляционного анализа.
3. Выполнить корреляционный анализ на массиве временных рядов.
4. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание набора временных рядов.
   3. Описание метода корреляционного анализа
   4. Описание результатов корреляционного анализа.
   5. Список использованных источников.
5. Оформить и защитить отчет.

**Контрольные вопросы**

1. Какие есть методы корреляционного анализа?
2. Каковы проблемы корреляционного анализа временных рядов.
3. К каким типа временных рядов возможно применение методов корреляционного анализа?
4. Какие практические задачи можно решать на основе корреляционного анализа временных рядов?
5. Каковы особенности у представленных на сайтах соревнований по анализу временных рядов?

## Лабораторная работа №4

Тема лабораторной работы: Прогнозирование на основе статистического подхода.

**Методические рекомендации и материалы**

Модели основаны на допущении о том, что основные факторы и тенденции прошлого периода сохранятся и на период прогноза, или что направление и изменение тенденций в рассматриваемой перспективе можно обосновать и учесть, т.е. предполагается большая инерционность систем. Модель временного ряда. Для прогнозирования будущих показателей имеющегося временного ряда необходимо построить модель ряда, которая наиболее полно отражает изменение исследуемого ряда. Существует множество моделей, описывающих различные стохастические процессы и имеющие различные формы представления, но среди них выделяют следующие, имеющие наибольшую ценность в практике:

* авторегрессионные модели (первого порядка, второго порядка) - (Autoregressive Moving Average, ARMA) ;
* модели скользящего среднего
* интегральные модели (Autoregressive Integrated Moving Average, ARIMA) .

В целом, статистический подход к анализу временных рядов заключается в выявлении и моделировании его детерминированных компонент на основе аддитивной (или мультипликативной) параметрической функциональной модели, приведении остатков к стационарному виду, при моделировании которых полученные ошибки удовлетворяли ограничениям модели.

Изучение временных рядов осуществляется при помощи вероятностно-статистических моделей. При этом выделяют характеристики временного ряда :

* математическое ожидание ;
* дисперсия ;
* автокорреляционая функция временного ряда: .

В зависимости от значений этих показателей временные ряды делятся на стационарные и нестационарные.

* Стационарные временные ряды можно выделить с помощью такого критерия как неизменность ранее перечисленных характеристик временного ряда: математическое ожидание и дисперсия являются постоянными величинами, автокорреляционная функция зависит только от разности . Иначе временной ряд является нестационарным.
* Нестационарные временные ряды. В терминах статистики в поведении временного ряда обычно выявляют две основные тенденции - тренд и периодические колебания. При анализе временных рядов стараются выделить тренд. Если затем вычесть его из исходных данных то остается колеблющийся ряд - случайные скачки, нерегулярности. Существенным понятием для тренда является гладкость.

Можно выделить следующие методы представления тренда:

* функциональная форма, полином низкой степени;
  + плюсы: простая форма представления;
  + недостатки:
    - трудно производить обновление,
    - сложно оценивать параметры функции,
    - подбор новой функции после добавления новых точек ряда может исказить полученную ранее последовательность.
* скользящие средние;
  + плюсы: упрощенная форма полиномиального представления;
  + минусы: возможность сгладить циклическую, краткосрочную составляющую тренда, запаздывание значений относительно исходного ряда;
* взвешенные скользящие средние;
  + плюсы: учитывается расстояние от точки до середины интервала сглаживания;
  + минусы: усиливает зависимость уровней ряда друг от друга;
* экспоненциальная средняя;
  + плюсы: влияние прошлых наблюдений затухает по мере удаления от момента, для которого определяется средняя;
  + минусы: влияние будет значительным для первых членов ряда, следовательно, ряд должен иметь достаточно большое количество уровней;
  + условия применения: наличие предыдущего значения в связи с рекуррентным процессом вычисления; выбор оптимальной постоянной сглаживания, характеризующей скорость реакции модели на изменение уровней.

Главный недостаток этих методов в том, что они рассматривают временной ряд изолированно от других явлений, и если даже имеется дополнительная информация, она может быть использована исследователем лишь путем регулирования скорости адаптации. Кроме того, точность прогнозов заметно падает при долгосрочном прогнозировании.

Следующим шагом после сглаживания тренда является моделирование стационарного процесса - разности между трендом и временным рядом. Такие ряды составляют другой класс временных рядов (в отличие от рядов, для которых возможно применить сглаживание с помощью скользящего среднего с конечным отрезком усреднения), известный как класс авторегрессий.

Минусы стационарных моделей: не существуют однозначных и эффективных критериев и методов определения факта наличия детерминированного тренда. Существуют статистические критерии проверки гипотезы о наличии тренда. Но эти критерии используют двухальтернативный базис: тренд или случайная компонента (метод восходящих-нисходящих серий), тренд или периодическая компонента, регулярная или случайная компонента.

Статистические модели характеризуются невысоким качеством при моделировании коротких временных рядов (количество наблюдений меньше 40)

Минусы статистических методов:

* отсутствие в модели представлений о структуре и системе связей реального объекта, что вносит субъективизм в выбор как самой модели, так и ее структуры;
* трудность построения моделей при условии, что данные хранятся в разных временных рядах и (или) имеют временные сдвиги относительно друг друга;
* значительная чувствительность получаемых результатов к недостатку информации и (или) ее зашумленности;
* потребность в высокой квалификации математиков-программистов;
* зависимость результата прогноза от квалификации аналитика в конкретной предметной области.

**Задания к лабораторной работе**

1. Для одного из соревнований по прогнозированию выбрать временные ряды из разных категорий.
2. Сформировать модели временных рядов и построить прогнозы.
3. Оценить качество прогнозов.
4. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание набора временных рядов.
   3. Описание моделей временных рядов.
   4. Описание прогнозирования и оценка качества.
   5. Список использованных источников.
5. Оформить и защитить отчет.

**Контрольные вопросы**

1. Постановка задачи, основные задачи анализа ВР. Критерии качества моделей. Стационарные и нестационарные временные ряды.
2. Какие основные классы методов анализа ВР? Data-driven и model‐driven методы анализа. Проблемы прогнозирования.
3. Принципы прогнозирования в статистическом подходе к анализу ВР.
4. Декомпозиция ВР, типы паттернов.
5. Модели тренда ВР (на основе функций от времени)

## Лабораторная работа №5

Тема лабораторной работы: прогнозирование на основе нейросетевого подхода.

**Методические рекомендации и материалы**

Гибкость и возможности нейронных сетей применительно к распознаванию сделали их привлекательной альтернативой, когда структура данных порождающей системы неизвестна. Однако, если начать формулировать прогностическую модель, ИНС, как правило, трудно интерпретировать и для проверки статистической значимости параметров.

Нейронные сети благодаря своим свойствам хорошо зарекомендовали себя при обработке данных. Одной из важных особенностей является способность к обучению и обобщению накопленных знаний. На ограниченном множестве данных сеть, обобщая полученную информацию, показывает хорошие результаты на данных, которые не использовались при обучении.

Функции, выполняемые сетями:

* аппроксимация,
* классификация и распознавание образов,
* прогнозирование,
* идентификация и оценивание,
* ассоциативное управление.

При прогнозировании роль нейронной сети состоит в предсказании будущей реакции системы по ее предшествующему поведению. Обладая информацией о значениях переменной в моменты, предшествующие прогнозированию 𝑥(𝑘 — 1),𝑥(𝑘 — 2),…,𝑥(𝑘 — 𝑁), сеть вырабатывает решение, каким будет наиболее вероятное значение последовательности 𝑥(𝑘) в текущий момент.

Моделирование временного ряда в рамках нейросетевого подхода состоит в аппроксимации нелинейной функции от многих переменных по набору примеров, заданных историей временного ряда:

где - прогнозируемое значение уровня временного ряда;

- наблюдаемые значения уровней временного ряда;

- некоторая нелинейная функция, параметрической моделью

которой служит нейронная сеть; ошибка прогноза;

n - порядок модели.

Эффективность нейронной сети во многом зависит от ее структуры. Взаимодействие между различными узлами сети задаются с помощью структуры. Структура ИНС не является уникальной для данной задачи, и могут существовать различные способы определить структуру, соответствующую проблеме. В зависимости от задачи, может оказаться целесообразным иметь больше одного скрытого слоя, упреждающие или обратные связи, а в некоторых случаях, прямые связи между входным и выходным слоем.

Успех нейросетевого моделирования зависит от

* типа данных;
* умения аналитика в выборе подходящей нейросетевой модели и / или;
* численных методов, используемых в модели и для вычисления прогнозов.

Исследования показали, что хорошая модель нейросетевая модель для временных рядов должна быть выбрана путем сочетания традиционного моделирования со знанием анализа временных рядов и проблем, связанных с параметрами нейросетевых моделей.

Влияние на качество нейросетевой модели оказывает также размер окна -количество данных, участвующих в прогнозе. Если окно слишком мало, то аттрактор системы проецируется на пространство недостаточной размерности. Кроме того, окно слишком большого размера может привести к проблемам: кроме основной информации в окно попадает шум.

Выделяется три сущности, необходимые для построения нейронной сети:

* Модель нейронной сети, архитектуры.
* Алгоритм обучения, который определяет веса связей.
* Функция, которая определяет выход каждого нейрона, функция активации.

Модели нейронной сети могут быть разделены на следующие три типа:

* Сети c прямой связью: рассматривают восприятие моделью обратного распространения, в основном используется в таких областях, как прогнозирование и распознавания образов;
* Сети с обратной связью: в основном используется для ассоциативной памяти и оптимизации расчетов;
* Самоорганизующиеся сети: рассматривают модели адаптивной резонансной теории и модели Кохонена, используется для кластерного анализа.

Выделяются преимущества нейросетевых моделей:

* высокая точность: нейронные сети могут аппроксимировать сложные нелинейные отображения;
* устойчивость к шуму: нейронные сети являются очень гибкими по отношению к неполным, пропавшим и зашумленным данным;
* независимость от предыдущих предположений: нейронные сети не делают априорных предположений о распределении данных или формы взаимодействия между факторами;
* простота обслуживания: нейронные сети могут быть использованы с новыми данными, что делает их полезными в динамических средах;
* могут быть реализованы в параллельном оборудовании;
* когда элемент нейронной сети выходит из строя, она может продолжать работать без каких-либо проблем.

Возникающие проблемы:

* нет общих методов для определения оптимального количества нейронов, необходимого для решения любой задачи;
* трудно выбрать набор обучающих данных, который будет достаточен для решения задачи.

Искусственные нейронные сети имеют как общие проблемы, такие как сходимость, устойчивость, наличие минимальных параметров настройки, так и частные проблемы анализа временных рядов: малая скорость обучения, попадание в локальный минимум, трудность определения параметров тренировки. Объединение генетических алгоритмов с нейронными сетями помогают достичь более высоких результатов.

**Задания к лабораторной работе**

1. Для одного из соревнований по прогнозированию выбрать временные ряды из разных категорий.
2. Сформировать структуру нейронной сети для моделирования временных рядов и построить прогнозы.
3. Оценить качество прогнозов.
4. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание набора временных рядов.
   3. Описание моделей временных рядов.
   4. Описание прогнозирования и оценка качества.
   5. Список использованных источников.
5. Оформить и защитить отчет.

**Контрольные вопросы**

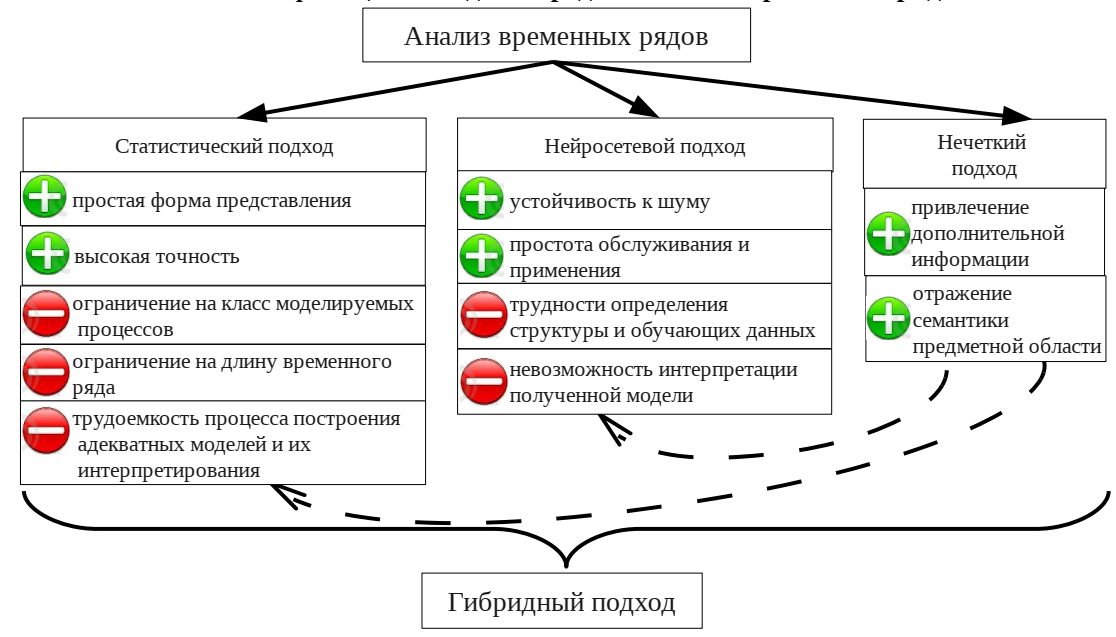
1. Каковы особенности архитектур нейронных сетей для моделирования временных рядов?
2. Как производится обучение нейронных сетей для моделирования временных рядов?
3. Как измеряется качество нейросетевых моделей временных рядов?
4. Каковы особенности применения нейронных сетей для моделирования временных рядов?
5. Какие задачи моделирования временных рядов могут быть решены на основе нейронных сетей?

## Лабораторная работа №6

Тема лабораторной работы: прогнозирование на основе нечеткого подхода.

**Методические рекомендации и материалы**

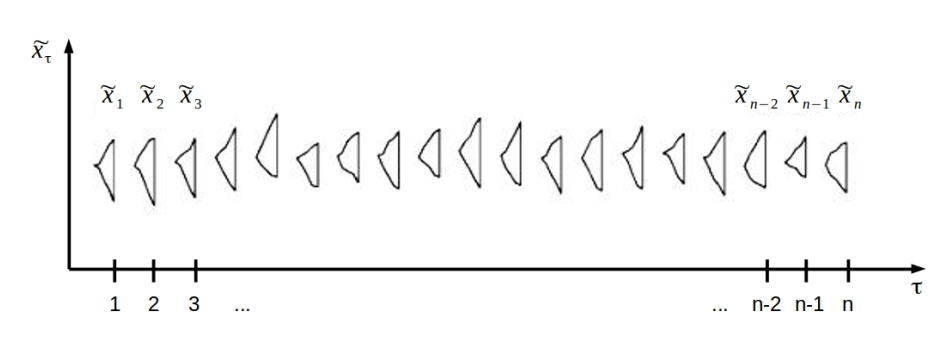
Использование моделей временных рядов в чистом виде в настоящее время может быть ограничено сложностью решаемых задач: сложно подобрать класс модели, невозможно добиться гибкости в отношении анализируемых данных, сами временные ряды могут иметь специфическое представление (быть нечеткие, обладать пропусками данных, аномалиями, и т. д.). Поэтому более востребованными и удобными в использовании в настоящее время являются различные гибридные модели, которые используют преимущества отдельных классов моделей, и устраняют недостатки за счет коллективной работы совокупности инструментов.



**Рисунок 1. Гибридизация моделей.**

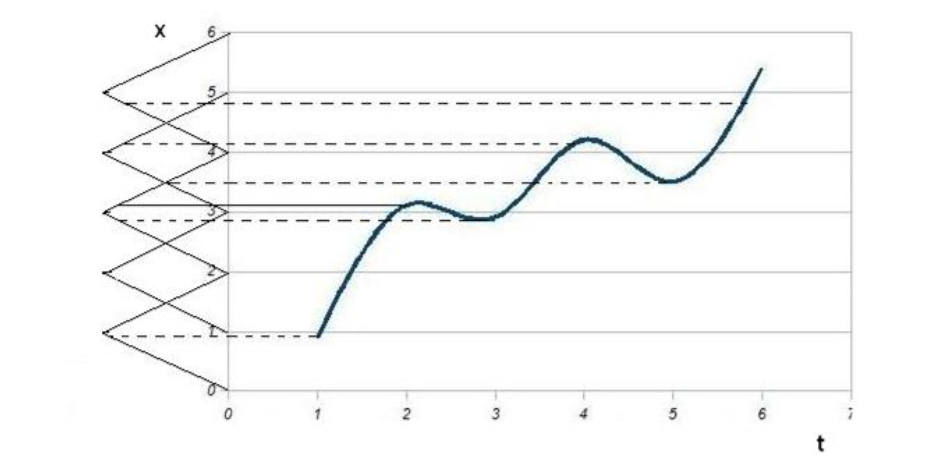
Под нечетким временным рядом (НВР) будем понимать упорядоченную последовательность наблюдений над некоторым явлением, состояния которого изменяются во времени, если значение состояния в момент выражено с помощью нечеткой метки- количество членов ряда, то есть нечеткий временной ряд представим в виде ,

где i-тое нечеткое множество (нечеткая метка), - i-тое значение момента времени n, - количество членов НВР.



**Рисунок 2. Нечеткий временной ряд.**

На рисунке 2 представлен нечеткий временной ряд. Визуализация выполнена в виде треугольных функций принадлежности. Изначально, временной ряд мог быть четким, необходимо выполнять фаззификацию. На рисунке 3 показана процедура, когда из четкого временного ряда можно получить нечеткий на основании расчета степени принадлежности в соответствующей точке.



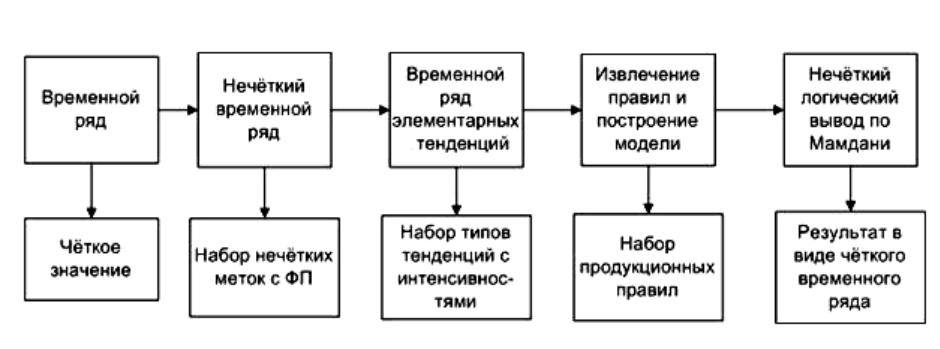
**Рисунок 3. Фаззификация временного ряда.**

Нечеткая метка — это понятие на естественном языке, получаемое посредством фаззификации исходного четкого временного ряда. Таким образом, для каждому четкому элементу временного ряда соответствует пара , где - лингвистическое понятие, - степень принадлежности элемента указанному понятию.

В отличие от традиционного временного ряда значениями нечеткого временного ряда являются нечеткие множества, а не действительные числа наблюдений. В 1993 году Сонг и Чиссом(Song, Chissom) предложили модели стационарных и нестационарных (time-invariant и time-variant) нечетких временных рядов первого порядка (first-order) и применили разработанные модели для прогнозирования количества регистрирующихся студентов университета штата Алабама, фаззифицировав предварительно четкий временной ряд. Это было первое определение моделей нечетких временных рядов.

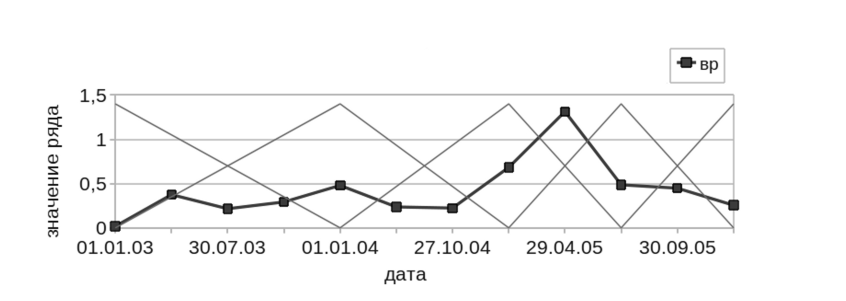
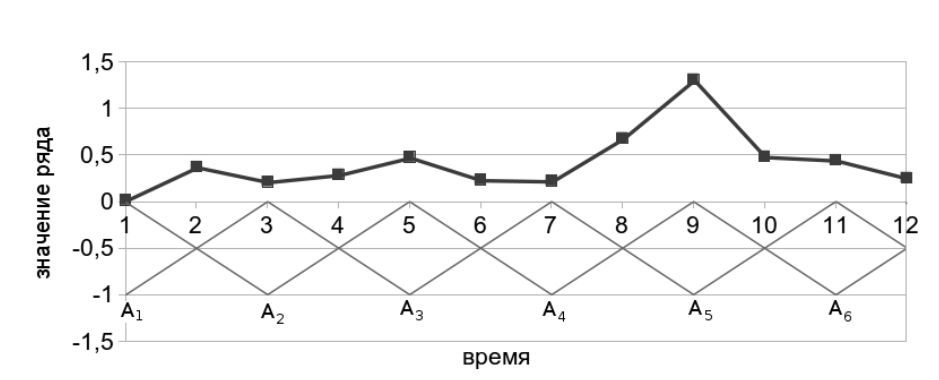
Пусть , - универсум, на котором определены нечеткие множества и - коллекция . Тогда называется нечетким временным рядом.

Все этапы анализа временного ряда с использованием нечеткого подхода представлены на рисунке 4.



**Рисунок 4. Принцип анализа временного ряда с использованием нечеткой логики.**

С точки зрения нечеткого подхода возможно введение нечеткости не только с точки зрения значений временного ряда, но и с точки зрения времени. На рисунке 5 показано как при помощи нечеткого представления времени можно группировать тенденции временного ряда.



**Рисунок 5. Введение нечеткости по времени.**

При использовании систем нечеткой логики необходимо определять систему нечеткого логического вывода. Следует описать основные компоненты системы нечеткого логического вывода:

* Входные переменные x, которые приводятся к нечетким значениям посредством вычисления их степени истинности выражения: есть , как вычисления значения соответствующего значения входной функции принадлежности, где u – номер входной переменной, j – номер правила нечёткого логического вывода;
* Треугольная функция принадлежности ;
* Операция пересечения нечётких множеств (t-норма);
* Лингвистические;
* Базовые терм-множества для лингвистических переменных;
* Выходные переменные y.

Разработка и применение систем нечеткого вывода включает в себя ряд этапов, которые выполняются последовательно, а именно:

1. Формирование базы правил

База правил — это множество правил, где каждому подзаключению сопоставлен определенный весовой коэффициент.

2. Фаззификация входных переменных

Целью этого этапа является получение значений истинности для всех подусловий из базы правил.

3. Агрегирование подусловий

Целью этого этапа является определение степени истинности условий для каждого правила системы нечеткого вывода.

4. Активизация подзаключений

На этом этапе происходит переход от условий к подзаключениям.

5. Акумуляция заключений

Целью этого этапа является получение нечеткого множества, или их объединения, для каждой из выходных переменных.

6. Дефаззификация выходных переменных

Цель дефаззификациии получить количественное значение для каждой из выходных лингвистических переменных.

Рассмотренные этапы нечеткого вывода могут быть реализованы неоднозначным образом: активизация может проводиться различными методами нечеткой композиции, дефаззификация также может проводиться различными методами. Таким образом, выбор конкретных способов реализации отдельных этапов нечеткого вывода определяет тот или иной алгоритм нечеткого вывода.

Существует и большое разнообразие методов и моделей для их представления, и что самое важное, их прогнозирования. В результате возникает задача выбора адекватного метода моделирования и прогнозирования динамического процесса или разработки нового гибкого адаптивного метода. Классификация моделей временных рядов строится на выделении трендовых и сезонных компонент временного ряда и построение на их основе моделей в форме сочетания аддитивной или мультипликативной зависимости. Сложность выбора конкретного метода для практического применения из большого количества существующих методов заключается в том, что не определены особенности временных рядов, которые однозначно определяют выбор метода прогнозирования. Временные ряды находятся в изменяющихся во времени условиях. Для решения проблемы является разработка массивов моделей и их объединения. В таком случае отмечается, что качество комбинированного прогноза часто выше качества составных прогнозов. Предлагается использовать модифицированного алгоритма агрегации. Другое решение – представление сведений об изучаемом объекте в виде иерархии, например, отделение глобальной, редко изменяющихся информации от локальной.

где и – адаптивные веса, и *, .*

*G -* это глобальная информация. Если существует система нечетких логических отношений *,* и – средние точки лингвистический переменных соответственно, то

*,*

В свою очередь *L* определяется как:

*,*

где и есть начальная и конечная точки интервала , и – средние точки лингвистический переменных и

Таким образом стало возможным осуществить реализацию 4-х методов: сочетания методов без трендовой компоненты, без сезонности и при их наличии.

Перечислим модели.

* Без тренда и сезонности:

*.*

Здесь подразумевается, что операция фаззификации заменяет собой сглаживание.

* Без тренда с аддитивной сезонной компонентой периода *p:*
* Аддитивный тренд без сезонной компоненты*:*
* Аддитивный тренд с аддитивной сезонной компонентой периода *p:*

Большой интерес представляют различные коллективы методов, использующие сведения о различных компонентах временных рядов для усиления качества модели. Если взять за основу гипотезу, что качество модели временного ряда отражается в качестве прогнозирования на ее основе, то необходимо описать информационные критерии, характеризующие такую модель. Выделяются следующие информационные критерии:

* Информационный критерий Акаике позволяет оценивать модели временных рядов.

где N - длина временного ряда, - значение временного ряда в i-тый момент времени, - результат прогнозирования модели в i-тый момент времени, K -количество параметров модели.

* Модифицированный критерий Акаике для рядов малой длины, у которых соотношение .
* Байесовский информационный критерий, который совпадает с информационным критерием Акаике, отличается он более строгим штрафом за сложность модели.

Основной идеей данного подхода к комбинации является вычисление весов для моделей на используя ИК. В проекте реализован способ расчёта. Для расчёта весов рассчитывается разница между наименьшим значением критерия и текущим, например критерий Акаике:

Далее рассчитывается функция правдоподобия:

После этого рассчитаем веса, нормализовав функцию правдоподобия:

Далее при расчёте прогноза будем использовать значения модели с учётов веса модели. Подобным образом можно рассчитывать веса, используя другой критерий, например MAPE, SMAPE.

**Задания к лабораторной работе**

1. Для одного из соревнований по прогнозированию выбрать временные ряды из разных категорий.
2. Построить нечеткие модели временных рядов и построить прогнозы.
3. Оценить качество прогнозов.
4. Составить отчет со следующей структурой:
   1. Цель работы.
   2. Описание набора временных рядов.
   3. Описание моделей временных рядов.
   4. Описание прогнозирования и оценка качества.
   5. Список использованных источников.
5. Оформить и защитить отчет.

**Контрольные вопросы**

1. Нечеткий подход к прогнозированию ВР. Этапы анализа и прогнозирования.
2. Методы прогнозирования ВР в нечетком подходе. Базовая модель нечеткого ВР Q. Song & B. Chissom (S-модель) и ее разновидности.
3. Виды моделей нечеткого логического вывода, применяемые при прогнозировании нечетких ВР (Мамдани, Суджено)
4. Задача анализа нечетких тенденций ВР. Формализация нечеткой тенденции. Виды нечетких тенденций
5. Основные задачи анализа ВР в терминах нечетких тенденций.
6. Возможности перехода к лингвистическим ВР в нечетком подходе.
7. Проблемы и преимущества прогнозирования ВР в нечетком подходе.
8. Критерии качества прогнозирования в решении прикладных задач.

## Список литературы

1. https://elibrary.ru/item.asp?id=13078384&
2. http://www.raai.org/about/persons/laritchev/papers/text/Larichev\_1981.pdf
3. Business Process Mining. Asef Pourmasoumi and Ebrahim Bagheri
4. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009
5. Business Process Mining: From Theory to Practice. Turner CJ, Tiwari A, Olaiya R, Xu Y
6. Data mining и big datа в бизнес-аналитике цифровой трансформации государственного и корпоративного управления. Тимофеев Александр Гурьевич, Лебединская Ольга Гурьевна
7. Process Mining: Overview and Opportunities WIL VAN DER AALST
8. Process Mining Manifesto
9. http://venec.ulstu.ru/lib/disk/2010/Yaruwkina.pdf
10. http://pzs.dstu.dp.ua/DataMining/times/bibl/afanasev.pdf
11. Андерсон, Т. Статистический анализ временных рядов / Т. Андерсон. — М: Мир, 1976. — Vol. 756.
12. Афанасьев, В.Н. Анализ временных рядов и прогнозирование / В.Н. Афанасьев, М.М. Юзбашев. — М: Финансы и статистика, 2001.— С. 228.
13. Кендэл, М. Временные ряды / М. Кендэл. — М: Финансы и статистика, 1981. — P. 199.
14. https://wiki.merionet.ru/servernye-resheniya/78/7-moshhnyx-baz-dannyx-vremennyx-ryadov
15. <https://habr.com/ru/company/oleg-bunin/blog/329062/>
16. https://temofeev.ru/info/articles/kak-my-testirovali-neskolko-baz-dannykh-vremennykh-ryadov

## Приложение 1. Временные ряды NN3

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **101** | **102** | **103** | **104** | **105** | **106** | **107** | **108** | **109** | **110** | **111** |
| 4998 | 5033 | 5304 | 685 | 5007 | 4950 | 4318 | 1740 | 6318 | 2605 | 3140 |
| 4480 | 4510 | 4264 | 656 | 5165 | 4900 | 4224 | 1005 | 5991 | 2510 | 3270 |
| 4824 | 3970 | 3224 | 1288 | 4976 | 5500 | 4014 | 1793 | 6405 | 3520 | 3690 |
| 4814 | 3378 | 2400 | 3914 | 4984 | 5550 | 3620 | 1965 | 6230 | 3200 | 4360 |
| 4602 | 2866 | 1968 | 6837 | 5105 | 5900 | 3512 | 2085 | 6111 | 2550 | 3520 |
| 4499 | 2316 | 1532 | 7104 | 5026 | 5150 | 3590 | 2440 | 5317 | 2640 | 4140 |
| 4594 | 1895 | 1000 | 7464 | 4848 | 5700 | 3700 | 2318 | 4893 | 2280 | 3440 |
| 4600 | 8402 | 55024 | 6308 | 4963 | 6450 | 3752 | 2313 | 4867 | 2620 | 3750 |
| 4507 | 8040 | 53804 | 5549 | 4796 | 4800 | 3866 | 2668 | 4785 | 1940 | 4140 |
| 4606 | 7534 | 42540 | 6365 | 5042 | 6300 | 3892 | 3655 | 5099 | 1415 | 5610 |
| 4503 | 7136 | 26572 | 6344 | 5028 | 5700 | 3838 | 2245 | 5142 | 1775 | 5350 |
| 4801 | 6467 | 15748 | 5597 | 4668 | 4900 | 3936 | 3000 | 5335 | 2320 | 4270 |
| 4564 | 5662 | 10372 | 1178 | 4646 | 4900 | 4200 | 2795 | 5450 | 2585 | 3160 |
| 4142 | 4896 | 8980 | 1581 | 4912 | 5500 | 4184 | 3493 | 5161 | 2885 | 2560 |
| 4818 | 4065 | 7212 | 1870 | 5041 | 5550 | 3708 | 3098 | 5516 | 2870 | 3410 |
| 4408 | 3296 | 7504 | 5230 | 5101 | 5200 | 3708 | 3665 | 4850 | 3010 | 3670 |
| 4496 | 2594 | 3848 | 8046 | 5143 | 5500 | 3760 | 6555 | 4250 | 2710 | 3670 |
| 4587 | 2007 | 2684 | 7997 | 4998 | 4950 | 3708 | 3040 | 4451 | 3070 | 3200 |
| 4656 | 1514 | 1000 | 9216 | 5053 | 5900 | 3846 | 8878 | 5442 | 2050 | 3700 |
| 4799 | 6645 | 43560 | 8077 | 5024 | 5700 | 3748 | 3280 | 5426 | 2315 | 3430 |
| 4652 | 6222 | 40320 | 6021 | 5299 | 5300 | 3836 | 3203 | 5329 | 2140 | 4130 |
| 4638 | 5474 | 28096 | 6399 | 4947 | 5550 | 3790 | 4593 | 5150 | 1950 | 4870 |
| 4650 | 5136 | 17804 | 6857 | 4872 | 4600 | 3968 | 2600 | 5030 | 3150 | 6310 |
| 5185 | 4631 | 10008 | 3446 | 4563 | 5650 | 3756 | 4803 | 4833 | 4510 | 4680 |
| 5208 | 4164 | 6136 | 1544 | 4580 | 4950 | 3760 | 3685 | 4437 | 3685 | 2380 |
| 4477 | 3601 | 4216 | 2031 | 4909 | 4850 | 3986 | 3688 | 4334 | 5125 | 3430 |
| 4976 | 2969 | 2424 | 2561 | 5002 | 5050 | 4096 | 2615 | 4165 | 5750 | 2950 |
| 4670 | 2504 | 1840 | 5166 | 5127 | 5500 | 4194 | 3070 | 3704 | 4580 | 3130 |
| 4842 | 2055 | 1112 | 7470 | 5060 | 5300 | 4254 | 2598 | 3616 | 4555 | 3650 |
| 4713 | 1609 | 324 | 6849 | 4947 | 5250 | 4130 | 2618 | 3476 | 3430 | 2980 |
| 4804 | 1298 | 100 | 6968 | 4810 | 6100 | 4044 | 2500 | 3118 | 2700 | 3720 |
| 4996 | 8485 | 58676 | 7253 | 4589 | 5200 | 3766 | 2685 | 3056 | 2010 | 2960 |
| 4574 | 8164 | 54768 | 6413 | 4408 | 6400 | 3930 | 2500 | 2922 | 1965 | 4300 |
| 4841 | 7814 | 43624 | 5854 | 4558 | 5700 | 3910 | 3693 | 2755 | 1185 | 5550 |
| 4688 | 7454 | 28680 | 5710 | 4509 | 5000 | 4024 | 1178 | 2691 | 2180 | 4380 |
| 4766 | 6889 | 15008 | 4926 | 4331 | 5400 | 4158 | 2923 | 2586 | 2705 | 3620 |
| 4994 | 6284 | 9028 | 1300 | 4770 | 4250 | 4018 | 2788 | 2810 | 3345 | 3350 |
| 4514 | 5712 | 7240 | 2200 | 4926 | 4950 | 3836 | 2090 | 2740 | 11760 | 3300 |
| 4766 | 5030 | 4884 | 2782 | 4679 | 5750 | 3764 | 2735 | 2565 | 3665 | 3800 |
| 4642 | 4488 | 3812 | 4708 | 4775 | 4600 | 3684 | 2855 | 3204 | 2490 | 3420 |
| 4806 | 4059 | 3696 | 6792 | 4707 | 4450 | 3668 | 3225 | 3149 | 2290 | 3860 |
| 4645 | 3585 | 2912 | 6193 | 4609 | 5400 | 3770 | 2218 | 2732 | 1600 | 2900 |
| 4784 | 3200 | 1200 | 5647 | 4644 | 4600 | 3888 | 2060 | 3267 | 1320 | 3240 |
| 4979 | 8181 | 42468 | 4317 | 4641 | 5400 | 3808 | 1965 | 3250 | 4950 | 3150 |
| 4530 | 8220 | 41900 | 4628 | 4436 | 4850 | 3794 | 1755 | 3194 | 1305 | 4460 |
| 4942 | 7866 | 30180 | 4658 | 4574 | 4100 | 3646 | 2465 | 3304 | 1310 | 5040 |
| 4651 | 7517 | 16836 | 4552 | 4503 | 4650 | 3712 | 2433 | 3332 | 2390 | 4660 |
| 5150 | 7116 | 9732 | 4349 | 4292 | 4700 | 3848 | 2465 | 3458 | 3685 | 3450 |
| 4987 | 6616 | 5728 | 1642 | 4479 | 3900 | 4064 | 2385 | 3787 | 3265 | 2660 |
| 4532 | 6217 | 5728 | 1069 | 3930 | 4200 | 4156 | 2458 | 3588 | 3770 | 2400 |
| 5046 | 5700 | 3584 | 2244 | 3860 | 5050 | 3724 | 2038 | 3839 | 4365 | 2200 |
| 4783 | 5179 | 3068 | 4551 | 4065 | 4250 | 3520 | 2538 | 4177 | 3380 | 3020 |
| 4958 | 4728 | 2992 | 5242 | 3741 | 4900 | 3796 | 2165 | 4097 | 2420 | 3450 |
| 4815 | 4225 | 1092 | 6225 | 3651 | 5250 | 3956 | 4390 | 3561 | 2490 | 4110 |
| 5055 | 3781 | 600 | 6175 | 3653 | 4250 | 3840 | 3045 | 3636 | 1980 | 2760 |
| 5152 | 7024 | 28268 | 6090 | 3532 | 5700 | 3794 | 2283 | 3618 | 1710 | 3390 |
| 4773 | 6558 | 26652 | 6431 | 3607 | 4950 | 3970 | 2838 | 3555 | 1755 | 3740 |
| 5147 | 6258 | 19068 | 6769 | 3778 | 4200 | 3740 | 3463 | 3315 | 1465 | 4830 |
| 4866 | 5863 | 10024 | 6393 | 4019 | 4450 | 3860 | 2410 | 3344 | 1910 | 4450 |
| 5311 | 5343 | 4636 | 6704 | 3884 | 4850 | 3928 | 3358 | 3470 | 1885 | 3140 |
| 5172 | 4756 | 4888 | 4632 | 3850 | 5600 | 4260 | 3678 | 3461 | 2190 | 4700 |
| 4734 | 4174 | 5100 | 2757 | 4042 | 5150 | 4180 | 2383 | 3280 | 1840 | 2330 |
| 5011 | 3452 | 4664 | 2999 | 4216 | 5650 | 4144 | 3128 | 3508 | 2435 | 2740 |
| 4957 | 2849 | 3848 | 6150 | 4117 | 4850 | 4058 | 2885 | 3849 | 3060 | 2480 |
| 4968 | 2351 | 3772 | 6780 | 4232 | 6000 | 3954 | 4093 | 3776 | 2320 | 2690 |
| 5049 | 1888 | 2964 | 6785 | 4212 | 5200 | 3890 | 7223 | 3281 | 4155 | 5850 |
| 5305 | 1417 | 1200 | 7229 | 4066 | 5400 | 3816 | 3398 | 3434 | 2045 | 2450 |
| 5067 | 7399 | 50368 | 7243 | 4148 | 5500 | 3800 | 3900 | 3417 | 1915 | 3720 |
| 5001 | 7013 | 47796 | 6907 | 4060 | 4700 | 3868 | 3085 | 3358 | 1610 | 2780 |
| 5252 | 6645 | 38376 | 6673 | 4192 | 4900 | 3682 | 2343 | 3163 | 1305 | 5690 |
| 4903 | 6239 | 17728 | 6348 | 4249 | 4550 | 3804 | 4533 | 3190 | 2205 | 4150 |
| 5408 | 5721 | 12832 | 7166 | 3991 | 4500 | 3822 | 3503 | 3311 | 3160 | 4460 |
| 5395 | 5138 | 7768 | 4162 | 4101 | 4950 | 4062 | 3193 | 3505 | 3315 | 2730 |
| 5150 | 4357 | 5192 | 3295 | 4216 | 4550 | 4102 | 3350 | 3322 | 3595 | 3580 |
| 5460 | 3751 | 3896 | 3903 | 4302 | 5350 | 3928 | 3495 | 3553 | 4480 | 2870 |
| 4968 | 3324 | 4128 | 5996 | 4253 | 4900 | 3842 | 4488 | 3599 | 3615 | 4210 |
| 5021 | 2861 | 3716 | 7331 | 4269 | 5050 | 3816 | 2868 | 3531 | 3035 | 4500 |
| 5118 | 2456 | 2500 | 6850 | 4193 | 5150 | 3534 | 4093 | 3069 | 2110 | 3560 |
| 5175 | 2028 | 1700 | 7178 | 4102 | 5300 | 3624 | 2878 | 2589 | 1825 | 3800 |
| 5420 | 8389 | 5304 | 7342 | 4250 | 5350 | 3708 | 3543 | 2576 | 2260 | 3260 |
| 5121 | 7910 | 45856 | 6699 | 4134 | 4950 | 3692 | 5508 | 2532 | 1320 | 3410 |
| 5450 | 7686 | 34592 | 6642 | 4288 | 5150 | 3646 | 3005 | 2519 | 1460 | 3860 |
| 5286 | 7163 | 19172 | 6176 | 4384 | 4800 | 3716 | 2775 | 2541 | 2420 | 5000 |
| 5693 | 6842 | 9496 | 5580 | 4193 | 4700 | 3840 | 2388 | 2637 | 3145 | 3840 |
| 5353 | 6449 | 5960 | 3401 | 4057 | 4450 | 4154 | 3008 | 2641 | 3965 | 3210 |
| 5017 | 6061 | 4500 | 2508 | 4235 | 4250 | 4118 | 3913 | 2502 | 3830 | 2930 |
| 5577 | 5739 | 7560 | 3128 | 4420 | 4900 | 3824 | 3383 | 2677 | 3125 | 3630 |
| 4987 | 5363 | 3608 | 5936 | 4462 | 4850 | 3804 | 4130 | 3294 | 2720 | 2760 |
| 5129 | 5081 | 2948 | 7095 | 4464 | 4850 | 3706 | 3433 | 3232 | 2130 | 3310 |
| 5249 | 4764 | 3148 | 7490 | 4330 | 5300 | 3602 | 3833 | 2809 | 1765 | 4320 |
| 5100 | 4523 | 600 | 8205 | 4413 | 5600 | 3842 | 4375 | 3263 | 1260 | 3010 |
| 5382 | 9057 | 4196 | 8163 | 4323 | 6150 | 3676 | 2413 | 3246 | 965 | 3440 |
| 5039 | 8352 | 32476 | 7080 | 4215 | 5500 | 3750 | 2760 | 3190 | 935 | 5100 |
| 5364 | 7683 | 21144 | 7521 | 4453 | 5550 | 3696 | 4285 | 3295 | 1035 | 4170 |
| 5193 | 7320 | 15652 | 6923 | 4484 | 4300 | 3742 | 3305 | 3323 | 1115 | 4540 |
| 5846 | 6708 | 10160 | 6347 | 4230 | 4800 | 3582 | 5885 | 3449 | 935 | 3410 |
| 5259 | 6205 | 7308 | 3169 | 4229 | 4650 | 3966 | 3490 | 3462 | 1980 | 2590 |
| 4809 | 5577 | 5568 | 1802 | 4363 | 4500 | 4060 | 2945 | 3294 | 900 | 2640 |
| 5297 | 4777 | 3724 | 2574 | 4442 | 4650 | 3748 | 2595 | 3592 | 880 | 2970 |
| 5034 | 4280 | 4064 | 5016 | 4452 | 4800 | 3720 | 2863 | 3505 | 815 | 3000 |
| 5243 | 3918 | 5024 | 6280 | 4455 | 5000 | 3604 | 2170 | 3682 | 380 | 2900 |
| 5150 | 3289 | 4388 | 6520 | 4355 | 5200 | 3548 | 2098 | 3350 | 275 | 3010 |
| 5296 | 2394 | 1200 | 6658 | 4409 | 5500 | 3626 | 2383 | 3601 | 105 | 3380 |
| 5596 | 8132 | 49504 | 6513 | 4387 | 5500 | 3484 | 5468 | 3414 | 1305 | 3060 |
| 4954 | 8121 | 43196 | 6964 | 4350 | 5950 | 3792 | 3538 | 3457 | 1730 | 3100 |
| 5250 | 7791 | 30608 | 6508 | 4448 | 5500 | 3638 | 3440 | 3545 | 1570 | 3940 |
| 5009 | 7412 | 19300 | 5969 | 4492 | 4800 | 3574 | 3458 | 3399 | 2645 | 5360 |
| 5113 | 6861 | 10100 | 6134 | 4205 | 5300 | 3648 | 2963 | 3564 | 15745 | 3280 |
| 5237 | 6197 | 5880 | 2851 | 4370 | 5000 | 3750 | 2943 | 3568 | 2145 | 3190 |
| 4575 | 5623 | 5248 | 2122 | 4489 | 4950 | 3610 | 2833 | 3295 | 2495 | 4950 |
| 5026 | 4856 | 3812 | 2791 | 4516 | 5250 | 3462 | 2970 | 3722 | 2975 | 2910 |
| 4842 | 4304 | 3164 | 5516 | 4537 | 4900 | 3418 | 3075 | 3613 | 2645 | 2820 |
| 5019 | 3854 | 3872 | 6887 | 4441 | 5300 | 3478 | 2490 | 3525 | 2560 | 3050 |
| 5063 | 3284 | 3044 | 7446 | 4447 | 5750 | 3574 | 3298 | 3368 | 2110 | 2930 |
| 5261 | 2862 | 1200 | 7158 | 4477 | 5400 | 3466 | 2538 | 3527 | 2000 | 2800 |
| 5327 | 9487 | 55896 |  | 4368 | 5750 | 3468 |  | 3336 | 1975 | 3220 |
| 5054 | 9061 | 49948 |  | 4396 | 5100 | 3586 |  | 3320 | 1475 | 3120 |
| 5269 | 8878 | 35900 |  | 4528 | 4550 | 3366 |  | 3599 | 1730 | 4000 |
| 5019 | 8558 | 19352 |  | 4554 | 4650 | 3500 |  | 3301 | 2900 | 4840 |
| 5315 | 8031 | 7828 |  | 4316 | 5000 | 3666 |  | 3195 | 3405 | 2630 |
| 5274 | 7405 | 4324 |  | 4482 | 4350 | 3858 |  | 3354 | 3130 | 2920 |
| 4899 | 6853 | 4104 |  | 4719 | 4250 | 3842 |  | 3025 | 3490 | 2350 |
| 5216 | 6175 | 3972 |  | 4855 | 5200 | 3668 |  | 3369 | 3675 | 2850 |
| 5029 | 5342 | 2828 |  | 4820 | 4600 | 3700 |  |  | 2705 | 2700 |
| 5110 | 4976 | 4304 |  | 4704 | 4800 | 3748 |  |  | 2440 | 2870 |
| 5093 | 4290 | 4876 |  | 4608 | 5100 | 3790 |  |  | 2615 | 2580 |

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ

Учебно-методическое пособие

Автор РОМАНОВ Антон Алексеевич

Усл. печ. л. 4

УлГТУ, 432027, г. Ульяновск, ул. Сев. Венец, д. 32.