МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ГРОЗНЕНСКИЙ ГОСУДАРСТВЕННЫЙ НЕФТЯНОЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
имени академика М.Д. Миллионщикова

Кафедра «Информационные технологии»

Бетербиева А.И.

Методические рекомендации к лабораторным работам по дисциплине «Модели и методы интеллектуального анализа данных»

для студентов, обучающихся по направлению подготовки
09.03.02 «Информационные системы и технологии»

Магистр

Грозный 2023

Составители:

Бетербиева А.И., ассистент кафедры «Информационные технологии»

Рецензент:

Методические указания предназначены для бакалавров по направлению подготовки 09.03.02 Информационные системы и технологии института прикладных информационных технологий.

Методические рекомендации рассмотрены и утверждены на заседании кафедры «Информационные технологии»

Протокол № _ от _____г

Рекомендовано к изданию редакционно-издательским советом ГГНТУ.

Содержание

Введение

В современном информационном обществе данные стали одним из наиболее ценных ресурсов, и их анализ и использование представляют существенный интерес для многих сфер деятельности. Интеллектуальный анализ данных является мощным инструментом для извлечения полезной информации, выявления скрытых закономерностей и паттернов, а также прогнозирования будущих событий, на основе имеющихся данных.

Целью данного методического пособия является обеспечение студентам необходимых знаний, навыков и практического опыта для успешного применения интеллектуальных моделей, и методов в анализе данных. Рекомендации предлагают структурированный подход, основанный на современных методах исследования данных, алгоритмах машинного обучения и искусственного интеллекта. Студенты ознакомятся с основными концепциями, получат практические навыки и узнают о лучших практиках в области интеллектуального анализа данных. Для достижения этой цели перед студентами ставятся следующие задачи:

1. Ознакомление с теоретическими основами: Методические рекомендации предоставляют студентам исчерпывающую информацию о современных моделях, методах и алгоритмах в области интеллектуального анализа данных. Они позволяют ознакомиться со всеми необходимыми концепциями и понять принципы их работы.

2. Практическое применение знаний: Рекомендации включают практические задания, примеры и кейс-стади, чтобы студенты могли применить полученные знания на практике. Это помогает усвоить материал более глубоко и развивает навыки работы с интеллектуальными моделями и методами.

3. Развитие аналитических навыков: Методические рекомендации помогают студентам развить аналитическое мышление и способность применять различные методы анализа данных для получения ценной информации. Они позволяют студентам научиться эффективно использовать

интеллектуальные модели и методы для решения разнообразных задач анализа данных.

4. Изучение современных подходов: Задачей методических рекомендаций также является ознакомление студентов с современными подходами и тенденциями в области интеллектуального анализа данных. Обучение с использованием новейших инструментов и технологий помогает студентам быть в курсе последних достижений и применять их в своей работе.

5. Формирование критического мышления: Методические рекомендации способствуют развитию способности студентов к критическому анализу и оценке результатов анализа данных. Решение сложных задач анализа требует умения анализировать и интерпретировать информацию, а также принимать обоснованные решения на основе полученных результатов.

Основная задача методических указаний – научить студентов использовать различные методы, такие как кластеризация, классификация, регрессия, ассоциативные правила, нейронные сети и другие, чтобы извлекать ценную информацию из структурированных и неструктурированных данных.

Лабораторная работа №1. Изучение опыта применения методов кластеризации данных

1.1. Цель работы и общие требования

Целью работы является изучение современных приложений методов кластеризации данных в области прикладной информатики и программной инженерии на примере зарубежного опыта и зарубежных публикаций.

Исходные данные: англоязычная статья.

Результаты должны быть представлены в виде текстового отчета, содержащего перевод статьи и краткую характеристику статьи.

Требование к отчету

1.Титульный лист

- С англоязычным названием и авторами, информацией, где и когда опубликовано, ссылка на статью;
- ФИО и группа;
- Используемые информационные технологии при выполнении работы.

2. Цель работы.

3. Краткое изложение на русском (не более 2-х страниц). Изложение должно содержать следующие вопросы и ответы на них.

1) Какую проблему и из какой области решают авторы в статье. Зачем нужно решить эту проблему?

2) Как решалась эта проблема раньше: на основе информации из статьи (должна быть таблица методов решения с указанием источников в квадратных скобках и недостатков)

3) Что предложили авторы нового в решении поставленной проблемы и для устранения какого недостатка (один абзац)?

4) С помощью каких известных методов, моделей и алгоритмов (кластеризации) решается поставленная проблема в статье? Какова схема (методика) решения?

5) Какие данные были использованы для проведения экспериментов? Источники данных, количество и характеристики, примеры.

6) Какие критерии качества и сравнения (сколько тестовых наборов) использованы в статье?

7) Каков итог решения проблемы и какие задачи требуется решать в будущем?

8) Какие недостатки приведенного в статье исследования Вы заметили?

4. Перевод двухколончатый: первая колонка – английский текст, вторая – перевод на русский. Рисунки не переводить, оставлять исходники. Термины, связанные с ПО, переводить и в скобках оставлять английское обозначение. Аббревиатуры и те термины, которые они обозначают, оставлять без изменения.

Все ссылки на источники оставлять.

5. Список литературы не переводить, оставлять без изменения.

1.2. Методические рекомендации и материалы

При выполнении лабораторной работы могут быть использованы информационные технологии поддержки работ по переводу с одного языка на другой, например, Google-переводчик, Яндекс-переводчик, Promt и др.

На первом этапе рекомендуется с помощью информации из лекции, методических материалов изучить назначение, особенности различных методов кластеризации данных.

Кластеризация (сегментация) – это разделение множества объектов на группы, обладающих схожими характеристиками. Методы кластеризации относятся к методам Data Mining – это автоматизированный процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей, то есть извлечения информации, которая может быть охарактеризована как знания.

Исходными данными для кластеризации являются числовые данные, представленные в табличной (матричной) форме.

С помощью кластеризации решаются задачи:

- Группировка многомерных данных;
- Объединение сходных объектов;
- Разделение объектов.

Каждый полученный кластер в результате кластеризации характеризуется следующими понятиями:

- Кластер имеет математические характеристики: центр, радиус, среднеквадратическое отклонение, размер кластера.
- Центр кластера – это среднее геометрическое место точек в пространстве переменных.
- Радиус кластера – максимальное расстояние точек от центра кластера.

Различают два вида иерархических методов:

- агломеративные методы, основанные на объединении объектов в группу;
- дивизимные методы при разделении объектов на группы. Типичным представителем итерационных методов является метод К-

средних. Основные этапы его реализации:

1. Первоначальное распределение объектов по кластерам. Выбирается число k и выбираются исходные центры кластеров.
2. Итеративный процесс.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т. е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

Существует множество методов кластеризации, некоторые приведены в таблице 1.

Таблица 1. Методы кластеризации

Критерии	Методы кластеризации числовых данных					Методы кластеризации категориальных данных	
	K-средних	Farthest first	EM	EM (мод.)	Метод ближайшего соседа	CLOPE	Large Item
Простота реализации	+	+	+	-	+	+	+
Относительно высокое быстродействие	+	+	-	-	-	+	-
Нетребовательность к объему памяти	+	+	+	+	-	+	+
Возможность выделения кластеров произвольной формы	-	-	-	-	+	-	-
Отсутствие необходимости задания количества кластеров	-	-	-	+	+	+	+
Работа с числовыми атрибутами	+	+	+	+	+	-	-
Работа с категориальными атрибутами	-	-	-	-	-	+	+
Индекс качества кластеры							

Для более подробного изучения методов кластеризации рекомендуется обратиться к следующим материалам: Чубукова И.А. Data_Mining http://lnfm1.sai.msu.ru/~rastor/Books/Chubukova-Data_Mining.pdf

1.3. Задания к лабораторной работе

Лабораторная работа выполняется по вариантам, представленным в таблице. Необходимо выполнить поиск и перевод статьи в соответствии с вариантом и указанными в разделе 1 требованиям.

Названия статей приведены в таблице 2.

1.4. Контрольные вопросы

1. В чем состоит цель кластеризации? Приведите формальную постановку задачи кластеризации.
2. Какую проблему и из какой области решают авторы в статье? Зачем нужно решить эту проблему?
3. Как решалась эта проблема раньше: на основе информации из статьи (должна быть таблица методов решения с указанием источников в квадратных скобках и недостатков)?
4. Что предложили авторы нового в решении поставленной проблемы и для устранения какого недостатка (один абзац)?
5. С помощью каких известных методов кластеризации решается поставленная проблема в статье? Какова схема проверки качества решения?

6. Какие данные были использованы для проведения экспериментов? Источники данных, количество и характеристики, примеры.
7. Каков итог решения проблемы и какие задачи требуется решать в будущем?
8. Какие недостатки приведенного в статье исследования Вы заметили?

Таблица 2. Названия англоязычных статей для перевода

№	**Название статьи**
1	Clustering Methodologies for Software Engineering
2	Assessing the State of Software in a Large Enterprise: A Twelve Year Retrospective
3	Component identification from existing object oriented system using Hierarchical clustering
4	Analyzing Software Measurement Data with Clustering Techniques
5	Analogy Based Software Project Effort Estimation Using Projects Clustering
6	Clustering and Classification of Software Component for Efficient Component Retrieval
7	Multiple Layer Clustering of Large Software Systems
8	Combining Clustering and Classification for Software Quality Evaluation
9	Towards identifying software project clusters with regard to defect prediction

.

Лабораторная работа №2. Программирование методов кластеризации данных

2.1. Цель работы и общие требования

В лабораторной работе необходимо изучить методы кластеризации, предложить модификацию выбранного метода и получить навыки в создании приложений для решения практической задачи анализа данных на основе методов кластеризации.

Исходные данные: определены в задании к лабораторной работе. *Результаты* должны быть представлены в виде текстового отчета и работающего ПО.

Отчет должен содержать

1. Титульный лист.
2. Цель, задание и требования.
3. Описание объектов исследования, входных данных (атрибутов объектов) и исследовательских вопросов.
4. Выбор метода кластеризации.
5. Формальную постановку задачи кластеризации, применительно к прикладной области, при этом требуется предложить улучшение выбранного метода.
6. Архитектуру разработанного ПО (IDEF0 (как есть/как должно быть), UML диаграммы) и используемые технологии.
7. Результаты кластеризации выбранным и модифицированным методами и их сравнительная оценка.
8. Заключение и выводы.
9. Список литературы и источников.

Требования к ПО

Программа должна обеспечивать ввод исходных данных, выполнение кластеризации выбранным методом, кластеризации модифицированным методом, вывода результатов кластеризации, то есть полученных кластеров в табличной форме, строки – наименование кластеров, столбцы – математические характеристики. Для каждого кластера – мощность (количество объектов), центр, среднее внутрикластерное расстояние.

2.2. Методические рекомендации и материалы

При выполнении лабораторной работы могут быть использованы информационные технологии поддержки создания приложений для анализа данных.

На первом этапе рекомендуется с помощью информации из лекции, методических материалов изучить назначение, особенности различных методов кластеризации данных. Полезно будет обратиться к методическим материалам предыдущей лабораторной работы и своему опыту.

Формально постановку задачи можно сформулировать следующим образом.

Представим исходные данные в виде базы многомерных данных MD (n × d), содержащей множество из n записей X1 , …, Xn, таких, что каждая запись Xi (i = 1,2,…,n) состоит из значений (x1i ,…, xdi), где d – количество атрибутов. Каждая запись содержит данные об одном объекте, каждый атрибут – это характеристика объекта.

Задача кластеризации (Data Clustering) состоит в том, чтоб для базы данных MD определить ее разбиение по строкам на множеств кластеров (групп) C1,…, Ck , так чтобы в каждом кластере содержались похожие («similar») строки в смысле значений атрибутов, а в разных – непохожие.

Основные этапы кластерного анализа включают:

1. Отбор выборки для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке. Выполнение нормализации данных.
3. Вычисление значений той или иной меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов.
5. Проверка достоверности результатов кластерного решения. Различают иерархические и итерационные методы кластеризации.

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

В кластеризации используются различные меры сходства:

- Евклидово расстояние;
- Метрика Махаланобиса;
- Расстояние Чебышева. Это расстояние стоит использовать, когда необходимо определить два объекта как «различные», если они отличаются по какому-то одному измерению;
- Процент несогласия для категориальных данных;
- Метрика Левенштейна для данных в виде слов, используется в поисковиках [https://habrahabr.ru/post/114997/].

После получений результатов кластерного анализа следует проверить правильность кластеризации (т. е. оценить, насколько кластеры отличаются друг от друга).

Для этого рассчитываются математические характеристики для каждого кластера.

При хорошей кластеризации должны быть получены сильно отличающиеся кластеры по их математическим характеристикам или по средним для всех атрибутов объектов, попавших в отдельный кластер.

Для более подробного изучения методов кластеризации рекомендуется обратиться к следующим материалам:

1) Чубукова И.А. Data_Mining http://lnfm1.sai.msu.ru/~rastor/Books/Chubukova-Data_Mining.pdf,

2) Воронина В. В. Теория и практика машинного обучения: учебное пособие / В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. – Ульяновск: УлГТУ, 2017. – 290 с.

При проектировании системы интеллектуального анализа рекомендуется изучить разделы в книгах:

1) Афанасьева, Т. В., Афанасьев А.Н. Введение в проектирование систем интеллектуального анализа данных: учебное пособие. – Ульяновск: УлГТУ, 2017. 64 с.

2.3. Задания к лабораторной работе

Содержательные требования к лабораторной работе

1) Сформировать структуру данных для анализа объекта исследования и заполнить ее. Разработать методику и программу для сегментации объекта исследования.

2) Провести сегментацию объекта исследования по количественным признакам на основе выбранного метода кластеризации внутри каждого кластера.

3) Провести пространственную (по странам, регионам, городам, предметным областям, корпорациям сегментацию объекта исследования.

4) Провести временную сегментацию на основе тенденций «рост», «падение», «стабильность».

5) Сформулировать выводы и объяснить результаты. Варианты заданий:

1. Кластеризация рынка технологий Big Data;
2. Кластеризация языков программирования;
3. Кластеризация рынка IT профессий;
4. Кластеризация рынка рекламных технологий (ТВ, радио, интернет);
5. Кластеризация рынка IT продуктов;
6. Кластеризация рынка IT технологий разработки ПО;
7. Кластеризация рынка e-learning в области ИТ;
8. Кластеризация рынка распределенных систем;
9. Кластеризация рынка IoT;
10. Кластеризация регионов по индексам развития информационного общества;
11. Кластеризация средств визуального моделирования;
12. Кластеризация технических текстов при разработке ПО;
13. Кластеризация программных проектов по метрикам качества;
14. Кластеризация абитуриентов ИТ-направлений вузов;
15. Кластеризация покупателей по типу поведения;
16. Кластеризация ИТ-разработчиков по типу поведения;
17. Кластеризация пользователей ИТ-продуктов;
18. Кластеризация отзывов на ИТ-продукт;
19. Своя тема.

2.4. Контрольные вопросы

1. В чем состоит цель кластеризации? Приведите формальную постановку задачи кластеризации.
2. Приведите перечень и особенности методов кластеризации.
3. Какие метрики применяют в кластеризации?
4. Приведите математические характеристики кластеров и меры качества результатов кластеризации.
5. Охарактеризуйте этапы кластерного анализа выбранного метода.
6. Чем Ваше решение отличается от стандартного кластерного анализа на основе выбранного метода?

7. Сформулируйте выводы и объясните полученные результаты лабораторной работы.

Лабораторная работа №3. Лингвистическое резюмирование результатов кластеризации данных

3.1. Цель работы и общие требования

Целью работы является изучение и получение навыков в создании лингвистического описания (резюмирования) результатов кластеризации.

Исходные данные: Разработанное ПО и полученные результаты кластеризации данных прикладной области.

Результаты должны быть представлены в виде текстового отчета и новой функции ПО автоматические лингвистические описания результатов кластеризации.

Результаты должны быть представлены в виде текстового документа, включающего титульный лист, содержание документа:

1. Цель работы
2. Краткое описание объектов исследования
3. Наименование и тип выбранных атрибутов-признаков
4. Исходные данные для лингвистического резюмирования: результаты кластеризации, полученные в лабораторной работе

№2 (для каждого кластера список включенных объектов и их количественные атрибуты-признаки)

5. Постановка задачи лингвистического резюмирования результатов кластеризации по признакам (по пространственным атрибутам и по тенденциям изменения)
6. Модель лингвистической шкалы (лингвистической переменной) для каждого типа лингвистического резюмирования (с указанием используемых источников)
7. Результаты сгенерированного отчета лингвистического резюмирования
8. Выводы
9. Список литературы и источников.
10. При выполнении работы требуется разработать лингвистическую шкалу для генерации лингвистических оценок.

3.2. Методические рекомендации и материалы

При выполнении лабораторной работы могут быть использованы информационные технологии поддержки создания приложений для анализа данных.

На первом этапе рекомендуется с помощью информации из лекции, методических материалов изучить назначение, особенности различных методов лингвистического описания данных. Полезно будет обратиться к методическим материалам предыдущей лабораторной работы и своему опыту.

Полезно изучить разделы, посвященные ACL-шкале и лингвистическому резюмированию в работах:

1) Афанасьева Т. В., Ярушкина Н. Г. Нечеткое моделирование временных рядов и анализ нечетких тенденций, 2009 (разделы 4, 5.9);

2) Ярушкина Н. Г., Афанасьева Т. В., Перфильева И. Г. Интеллектуальный анализ временных рядов (учебное пособие), 2010;

3) Ярушкина Н. Г., Афанасьева Т. В., Перфильева И. Г. Интеллектуальный анализ временных рядов: учебное пособие – М. : ИД «ФОРУМ» ИНФРА-М, 2012 (разделы 3.4, пример 3.1).

При выполнении лабораторной работы предварительно нужно, проанализировав значения атрибутов объектов и мощность полученных кластеров, создать лингвистические шкалы (лингвистические переменные). Для этого рекомендуется обратиться к теории нечетких множеств. Каждая лингвистическая переменная будет связывать числовые характеристики кластеров с лингвистическими термами, название которых необходимо задать. Можно использовать названия лингвистических термов из примера, рассмотренного ниже.

Пример лингвистического описания кластеров:

Общая характеристика кластеров.

Общее количество исследуемых объектов равно «Столько-то», они были сгруппированы в «столько-то» кластеров: «Имя кластера 1»…, «Имя кластера К».

Количество объектов можно изменять, количество кластеров изменять («можно» или «нельзя»).

Большинство «объектов исследования» сгруппировано в кластере «Имя кластера».

Наименьшее количество «объектов исследования» сгруппировано в кластере «Имя кластера».

Или «объекты исследования» сгруппированы в кластеры **равномерно** (нужно предварительно определить лингвистические оценки «Большинство», «Меньшинство», «Равномерно»).

Характеристика каждого кластера может быть представлена в виде

ID_Кластера «Имя кластера» Содержит «Столько-то»процентов исследуемых объектов.

Характеристика признаков (по которым проводилась кластеризация) в кластере «Имя кластера»

1. «наименование признака 1»
   - Среднее значение = значение,
   - Минимальное значение = значение,
   - Максимальное значение=значение.
2. «наименование признака 2»
   - Среднее значение = значение,
   - Минимальное значение = значение,
   - Максимальное значение=значение.

Темпоральная характеристика «Имя кластера» по типам тенденции (типы тенденции: рост, падение, стабильность) формируется при наличии данных за несколько исторических периодов.

Типичная тенденция «тип тенденции» соответствует «наименование признака 1» и составляет «столько» процентов.

Нетипичная тенденция «тип тенденции» соответствует «наименование признака К» и составляет «столько» процентов.

Пространственное распределение в кластере «Имя кластера» при наличии географического распределения анализируемых данных.

Большинство «объектов исследования» относится к «наименование географического места» (или другого пространства), что составляет

«столько-то процентов».

Наименьшее количество «объектов исследования» относится к «наименование географического места» (или другого пространства), что составляет «столько-то процентов».

Или «объекты исследования» равномерно распределены в пространстве «наименование пространства».

Пример лингвистической шкалы для лингвистического описания атрибута доли аудитории для каналов распространения рекламы приведен в таблице 3.1.

Таблица 3.1. Пример лингвистической шкалы

Значение лингвистической переменной	**Значение доли аудитории (%)**
Большая часть	trapmf(65,70,100,100)
Больше половины	trimf(50,60,70)
Половина	trimf(45,50,55)
Меньшая половины	trimf(30,40,50)
Меньшая часть	trapmf(0,0,30,35)

Здесь использованы трапецеидальная и треугольные функции принадлежности для нечетких множеств.

Графические представления таких функций приведены на рисунке 3.1.

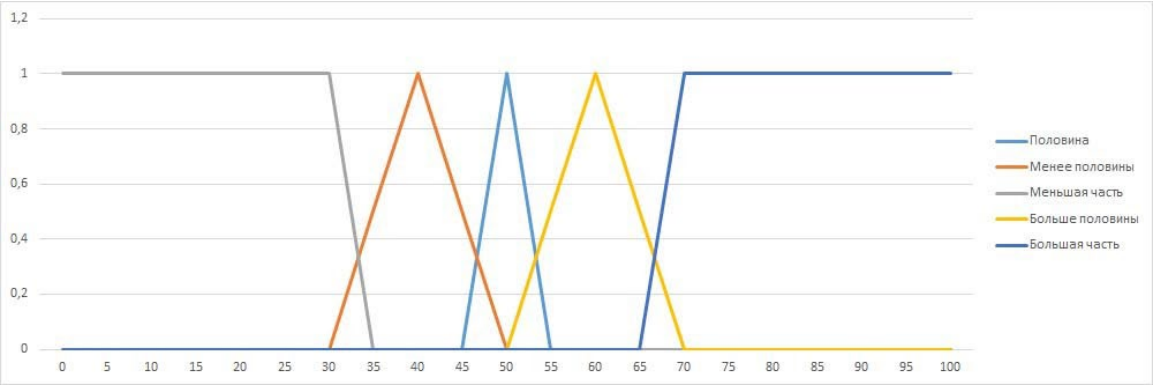

Рис. 3.1. Функции принадлежности для признакового резюмирования

Для оценивания тенденций изменения в качестве примера можно использовать следующую лингвистическую шкалу (таблица 3.2).

Таблица 3.2. Пример лингвистической шкалы

Значение лингвистической переменной	**Значение разницы доли аудитории между текущим и предыдущим периодом (%)**
Резкий рост	trapmf(30,35,100,100)
Значительный рост	trimf(15,20,31)
Уверенный рост	trimf(10,13.5,16)
Рост	trimf(5,7.5,11)
Слабый рост	trimf(1.5,2.5,6)
Стабильность	trimf(-2,0,2)
Слабое снижение	trimf(-6,-2.5,-1.5)
Снижение	trimf(-11,-7.5,-5)
Уверенное снижение	trimf(-16,-13.5,-10)
Значительное снижение	trimf(-31,-20,-15)
Резкое снижение	trapmf(-100,-100,-30,-35)

На рисунке 3.2 изображены графические представления нечетких множеств для темпорального резюмирования.

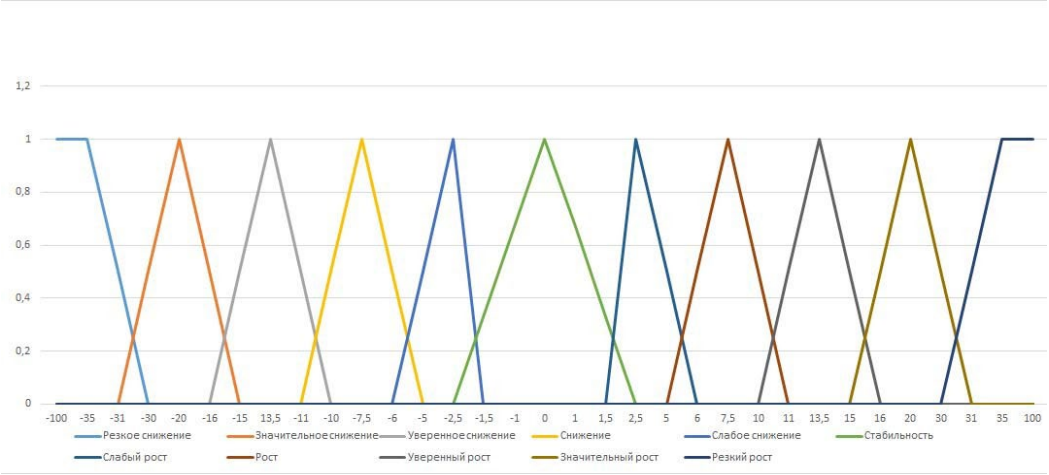

Рис. 3.2. Функции принадлежности для темпорального резюмирования

Полученные результаты применения разработанных лингвистических шкал представлены ниже.

Признаковая кластеризация.

Кластер «Самые популярные по аудитории»

- Максимальное значение*:* 1 375 100 072 чел.
- Минимальное значение*:* 1 106 898 836 чел.
- Среднее значение*:* 1 252 635 550 чел.
- Сумма по всей аудитории: 3 757 906 652 чел.
- Объем аудитории*:* Меньше половины
- Мощность кластера*:* Меньшая часть

3.3. Задание к лабораторной работе

Содержательные требования к лабораторной работе

Создать функцию автоматического лингвистического описания результатов кластеризации данных из прикладной области, выполненной в рамках лабораторной работы № 2.

С помощью модифицированного ПО выполнить лингвистическое описание кластеров и сформировать электронный и текстовый варианты отчета о кластеризации объектов прикладной области и лингвистического описания полученных кластеров.

Сформулировать выводы и объяснить результаты.

3.4. Контрольные вопросы

1. В чем состоит цель лингвистического описания данных? Приведите формальную постановку этой задачи.
2. Приведите определение лингвистической шкалы.
3. Приведите этапы создания лингвистической шкалы.
4. Сформулируйте выводы и объясните полученные результаты лабораторной работы.

Лабораторная работа №4. Подготовка научной статьи по результатам лабораторных работ № 1-3

4.1. Цель работы и общие требования

Цель работы состоит в получении навыков научного описания полученных результатов в виде научной статьи.

Исходные данные: Разработанное ПО и отчеты по лабораторным работам 1-3.

Результаты должны быть представлены в виде текста статьи на русском языке (Оригинальность – не менее 75%).

4.2. Методические рекомендации и материалы

При написании статьи рекомендуется предварительно изучить методику изложения научных результатов, например, используя структуру статьи из лабораторной работы № 1, а также учебные издания:

1) Семушин И. В. Письменная и устная научная коммуникация:
учебное пособие, 2014;

2) Афанасьева Т. В. Организация магистерских научно- исследовательских работ: методические рекомендации, 2015.

При описании в статье архитектуры приложения можно воспользоваться примером описания ПО, представленном в книге *Афанасьева Т. В., Ярушкина Н. Г. Нечеткое моделирование и анализ нечетких тенденций, 2009*.

Требования к оформлению статьи.

Статьи должны быть выполнены в текстовом редакторе MS Word 2003-2016 и отредактированы строго по следующим параметрам:

- ориентация листа – книжная,
- формат А4,
- поля по 2 см по периметру страницы,
- гарнитура Times New Roman,
- размер шрифта для всей статьи, кроме таблиц – 14 пт,
- размер шрифта для таблиц – 12 пт,
- междустрочный интервал – 1.5,
- выравнивание по ширине страницы,
- абзацный отступ – 1 см (без использования клавиш «Tab» или «Пробел»).

Не допускается:

- нумерация страниц;
- использование в тексте разрывов страниц;
- использование автоматических постраничных ссылок;
- использование автоматических переносов;
- использование разреженного или уплотненного межбуквенного интервала.

Таблицы набираются в редакторе MS Word. Таблицы должны иметь номера и названия, которые указывают над таблицами.

Графический материал (рисунки, схемы) должен представлять собой обобщенные материалы исследований. Графический материал должен быть высокого качества, названия и номера графического материала указывать под изображением.

Формулы и математические символы выполнять либо в MS Word с использованием встроенного редактора формул, либо в редакторе MathType.

Таблицы, графический материал и формулы не должны выходить за пределы указанных полей.

4.3. Методика выполнения лабораторной работы

При выполнении лабораторной работы могут быть использованы информационные технологии поддержки поиска аналогов.

На первом этапе рекомендуется сформулировать несколько заглавий статьи. Используя эти заглавия выполнить поиск аналогов в интернете (русскоязычном и англоязычном).

Найденные статьи сохранить. Затем выполнить обзор найденных статей (это раздел Похожие работы), кратко описав каждую статью, сфокусировав внимание на том, чем цель, постановка, методы или результаты отличаются от ваших. Это позволит определить традиционные подходы и выбрать аналогичную работу, относительно которой будут сравниваться ваши результаты.

Дальнейшая работа связана с описанием вашего решения и демонстрации его эффективности, которое доказывается путем сравнения с аналогом.

Параллельно следует оформлять список используемой литературы.

Заключительным этапом подготовки русскоязычной статьи будет ее оформление в соответствии с требованиями.

4.4. Задания к лабораторной работе *Содержательные требования к лабораторной работе*

Статья должна иметь следующие разделы.

1) Заглавие, отражающее основной результат.
2) ФИО автора.
3) Место работы или учебы.
4) E-mail.
5) Аннотацию (80-120 слов), кратко описывающую содержание статьи.
6) Введение, в котором кратко описывается решаемая проблема, кратко, какие подходы применяют для ее решения, обосновывается полезность улучшения традиционных подходов за счет Вашего решения.
7) Похожие работы. Эти публикации необходимо найти в англоязычном секторе интернета, а также использовать результаты работы № 1. Обязательны ссылки типа. Привести, чем отличается предлагаемое в вашей статье решение проблемы.
8) Описать постановку проблемы и предложенное решение по шагам с указанием особенностей метода кластеризации.
9) Описать архитектуру ПО и используемые технологии. Использовать схемы.
10) Описать контрольный пример: что использовали, сколько объектов на входе, какие признаки выбрали, сколько и какие кластеры получили. Привести скриншоты,

показывающие конечные результаты. Привести доводы или сравнительное исследование по некоторому критерию между вашим решением и выбранным аналогом.

11) Сформулировать выводы и объяснить результаты.

12) Список источников, не менее 7.

4.5. Контрольные вопросы

1) Приведите структуру научной статьи.

2) Приведите этапы написания научной статьи.

3) Опишите анализируемые методы кластеризации и принципы выбора аналогичного решения.

4) Сформулируйте ограничения выбранного аналога.

5) Охарактеризуйте, что нового в полученных результатах по сравнению с аналогом.

6) Сформулируйте выводы и объясните полученные результаты лабораторной работы.

Лабораторная работа №5. Прогнозирование на основе статистического подхода

5.1. Цель работы и общие требования

Цель работы: изучить задачу прогнозирования временных рядов на примере применения статистических моделей.

5.2. Методические рекомендации и материалы

Исходные данные: базы данных временных рядов (ВР) находятся: на

CIF_2015, CIF_2016 в разделе «Download» (http://irafm.osu.cz/cif/main.php):

1. База данных ВР, которая открыта для моделей Competition dataset(Train)
2. База данных ВР, которая закрыта для моделей Testing dataset (Test). Это фактически продолжения временных рядов, представленных в базе данных Train, поэтому количество ВР в базе данных Train равно количеству ВР в базе данных Test.

Используемое ПО:

a. Прогнозирование и декомпозиция ВР http://timeseries.greamko.ru/. Приложение разработано на языке R с использованием фреймворка Shiny. Отсутствует необходимость в регистрации, можно спрогнозировать тестовые ВР из файла формата «.csv».

b. Прогнозирование ВР статистическими моделями в системе http://forecast.greamko.ru. Для доступа к функционалу прогнозирования необходимо пройти авторизацию (логин – «tv.afanasjeva@gmail.com», пароль – «ПИмд21_2016»), сформировать и загрузить временной ряд из файла формата «.csv».

Пример (http://joxi.ru/Dr8vE99FkBXnR2.jpg). После авторизации на странице Прогнозирование можно будет выбрать загруженный ранее ВР для прогноза. Есть возможность сохранить результаты в файл PDF.

Перед выполнением лабораторной работы рекомендуется изучить лекцию по теме работы, составить представление о типах поведения и моделях временных рядов, этапах подбора и оценивания моделей.

Чтобы сформировать компетенцию по интеллектуальному анализу процессов в рамках статистического подхода к анализу временных рядов, рекомендуется изучить одноименные разделы в книгах:

1) Афанасьева Т. В., Ярушкина Н. Г. Нечеткое моделирование временных рядов и анализ нечетких тенденций, 2009. (Раздел 3.1)
2) Ярушкина Н. Г., Афанасьева Т. В., Перфильева И. Г. Интеллектуальный анализ временных рядов (учебное пособие), 2010.

Временной ряд (ВР) – это последовательность дискретных упорядоченных в неслучайные равноотстоящие моменты времени измерений (показателей, наблюдений) y(t1), y(t2), …, y(tN), характеризующих уровни состояний изучаемого процесса, протекающего в условиях неопределенности.

Пусть заданы значения временного ряда Y = {y(1), y(2),..., y(N)} , где y(t) – значение показателя исследуемого процесса, зарегистрированного в t-м такте времени (t = 1, 2, ..., N). Требуется построить оценки будущих значений ряда $Y^{\wedge} = \{y^{\wedge}(N + 1), y^{\wedge}(N + 2),..., y^{\wedge}(N + t)\}$, $1 £ t £ N$, , где τ – горизонт прогнозирования.

На рис. 5.1 изображена общая схема идентификации модели временного ряда.

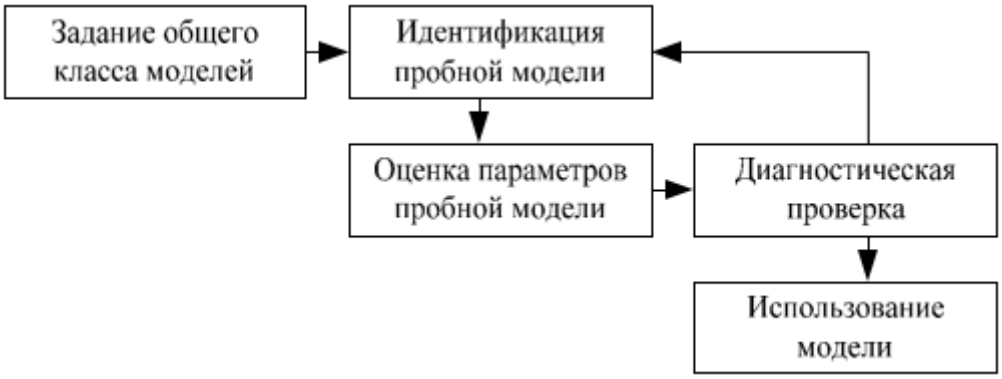

Рис. 5.1. Схема идентификации модели

Общей статистической моделью числового временного ряда служит модель вида: yt = f(xt,a) +εt .

В этой модели наблюдаемый ряд yt рассматривается как сумма некоторой систематической компоненты f(xt,a),где a – параметр, и случайной компоненты εt, рассматриваемой как независимые реализации случайного процесса типа «белый шум» с постоянным математическим ожиданием, постоянной и малой дисперсией.

Различают стационарный и нестационарный характер поведения временного ряда. Стационарный временной ряд отличается от нестационарного следующими свойствами: его математическое ожидание, дисперсия и ковариация не зависят от момента времени, в котором они вычисляются.

В качестве модели стационарных временных рядов используются модели ARIMA(p,d,q), в которых параметры структуры p, d, q, определяющие порядок модели, могут принимать нулевые значения:

1. Модель авторегрессии AR(p) связывает текущие значения временного ряда с прошлыми значениями и соответствует модели ARIMA(p,0,0). Формально модель авторегрессии AR(p) записывается в виде взвешенной суммы:

$$X(t) = f_0 + f_1 * X(t-1) + f_2 * X(t-2) + \ldots + f_p * X(t-p) + E(t),$$

где $X(t)$ – текущее значение уровня ряда в момент времени t;

$f_0, f_1, f_2, \ldots, f_p$ – оцениваемые параметры;

p – порядок авторегрессии;

Et – ошибка от влияния переменных, которые не учитываются в данной модели.

Задача заключается в том, чтобы оценить параметры f0 , f1 , f 2 , □, f p .

Их можно оценить различными способами, например, через систему уравнений Юла-Уолкера, для составления этой системы потребуется расчет значений автокорреляционной функции, или методом наименьших квадратов.

2. Модель скользящего среднего MA(q) связывает текущие значения уровня ряда со значениями предыдущих ошибок и соответствует модели ARIMA(0,0,q). Формально модель MA(q) представима в виде взвешенной суммы:

$$Z(t) = m + \varepsilon(t) - w1 * \varepsilon(t-1) - w2 * \varepsilon(t-2) - \ldots - wq * \varepsilon(t-q),$$

где $Z(t)$ – текущее значение уровня ряда в момент времени t;

m – константа, определяющая математическое ожидание временного ряда; $w_0, w_1, w_2, \ldots, w_q$ – оцениваемые параметры.

3. Комбинированные модели стационарных временных рядов ARMA(p,q) соответствуют модели ARIMA(p,0,q) и представляют собой объединение моделей AR(p) и MA(q).

Нестационарные временные ряды, приводящиеся к стационарным удалением тренда (или «взятием разности»), описываются моделью ARIMA(p,d,q), где параметр d указывает количество вычислений разности соседних уровней ВР.

Наиболее распространенные критерии точности моделирования и прогнозирования временных рядов представлены в таблице 5.1.

Таблица 5.1. Критерии точности моделей временных рядов

Критерий	**Формула расчета**
Средняя квадратичная ошибка (СКО)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Квадратный корень из средней квадратичной ошибки	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Средняя относительная ошибка	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert \cdot 100\,\%$
Симметричная средняя относительная ошибка	$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{(y_i + \hat{y}_i)/2}\right\rvert \cdot 100\,\%$

5.3. Задание к лабораторной работе

Для временных рядов согласно варианту (см. табл. 5.3) выполнить следующие этапы анализа ВР.

1. Познакомиться с возможностями прогнозирования ВР с помощью систем:

1.1. http://timeseries.greamko.ru/

1.2. http://forecast.greamko.ru

2. Провести декомпозицию заданных ВР и анализ на наличие паттернов тренда (T), сезонности (S) и случайного шума (R) временных рядов в системе: http://timeseries.greamko.ru/. Результаты занести в таблицу.

3. Провести прогнозирование и выбрать лучшую статистическую модель для заданных ВР по критерию минимума MAPE (test) в системе: http://forecast.greamko.ru. Результаты прогнозирования занести в таблицу 5.2.

Таблица 5.2. Пример заполнения результатов анализа и прогнозирования ВР

Id ВР	Длина	Горизонт	T	S	R	C_модель	Mape (train)	Mape (test)	F-модель	SMAPE
CIF-2015 -ts1	128	12	growth	12	C	ARIMA (1,1,0) (0,1,0)	3.23	5.34	D(1,0)	0.19

Пояснения к таблице 5.2:

1. Id ВР – это обозначение ВР, например, CIF-2015-ts1. Это значит временной ряд ts1 из базы данных CIF-2015.
2. Длина – это количество точек ВР.
3. Горизонт – количество точек для прогнозирования.
4. T – это тип паттерна тренда, может принимать значения из множества {growth или fall или no}.
5. S – это тип паттерна сезонности, задается периодом сезонности, например, 0 или 4 или 6 или 12 и т. д.
6. R – это тип паттерна случайной компоненты, задается видом случайных флуктуаций (C – стационарные или N – нестационарные).
7. C-Модель – обозначение статистической модели над графиком в системе http://timeseries.greamko.ru/.
8. Mape – критерии качества C-модели, задаются значением отдельно для обучающей (MAPE (train)) и тестовой части MAPE (test) ВР.
9. F-модель – это лучшая нечеткая модель в системе http://salx.pw/IFSA, может быть типа S (Song&Chissom, 1993), D (Hwang, 2004), T (Afanasjeva, 2012). Необходимо, чтобы каждая модель сопровождалась параметрами: порядок модели и тестовый отрезок, например, S(1,3), что означает модель типа S, порядок этой модели 1, тестовый отрезок 3. Эти параметры должны соответствовать одной, самой точной и адекватной нечеткой модели. Ее поведение в прогнозе должно соответствовать поведению прогноза, полученному ранее с помощью статистической модели в системе http://timeseries.greamko.ru/.
10. SMAPE – это внешний SMAPE выбранной нечеткой модели в системе http://salx.pw/IFSA.

Форма представления результатов

Лабораторная работа выполняется по вариантам (см. табл. 5.3). Временные ряды находятся в базе данных CIF: http://irafm.osu.cz/cif/main.php.

Результаты необходимо представить в виде двух документов. Первый документ – это электронная таблица (в формате XLSX). Второй документ (в формате DOCX), содержащий титул, задание, графики прогнозированных ВР, критерии точности.

5.4. Контрольные вопросы

При защите лабораторной работы необходимо ответить на три вопроса из списка контрольных вопросов:

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).

6. Модели случайной компоненты ВР (AR, MA, ARMA, ARIMA).

7. Модели сезонных колебаний (индексные методы адаптивные методы EST, спектральные методы, сезонная ARIMA).

Таблица 5.3. Варианты задания к лабораторной работе

Номер варианта	Id BP	Номер варианта	Id BP	Номер варианта	Id BP
1	CIF-2016-ts1, ts2, ts3	10	CIF-2016-ts28, ts29, ts30	19	CIF-2016-ts55, ts56, ts57
2	CIF-2016-ts4, ts5, ts6	11	CIF-2016-ts31, ts32, ts33	20	CIF-2016-ts58, ts59, ts60
3	CIF-2016-ts7, ts8, ts9	12	CIF-2016-ts34, ts35, ts36	21	CIF-2016-ts61, ts62, ts63
4	CIF-2016-ts10, ts11, ts12	13	CIF-2016-ts37, ts38, ts39	22	CIF-2016-ts64, ts65, ts66
5	CIF-2016-ts13, ts14, ts15	14	CIF-2016-ts40, ts41, ts42	23	CIF-2016-ts67, ts68, ts69
6	CIF-2016-ts16, ts17, ts18	15	CIF-2016-ts43, ts44, ts45	24	CIF-2016-ts70, ts71, ts72
7	CIF-2016-ts19, ts20, ts21	16	CIF-2016-ts46, ts47, ts48	25	CIF-2015-ts1, ts2, ts3
8	CIF-2016-ts22, ts23, ts24	17	CIF-2016-ts49, ts50, ts51	26	CIF-2015-ts4, ts5, ts7
9	CIF-2016-ts25, ts26, ts27	18	CIF-2016-ts52, ts53, ts54	27	CIF-2016-ts8, ts9, ts10

Лабораторная работа № 6. Прогнозирование на основе нечеткого подхода

6.1. Цель работы и общие требования

Цель работы. Изучить задачу прогнозирования временных рядов (ВР) на примере применения нечетких моделей для тех же временных рядов, которые использовались в лабораторной работе № 5.

6.2. Методические рекомендации и материалы

Используемое ПО

Прогнозирование ВР нечеткими моделями в системе http://salx.pw/IFSA. В этой системе все ВР уже загружены, регистрации не требуется.

Чтобы сформировать компетенцию по интеллектуальному анализу процессов в рамках нечеткого подхода к анализу временных рядов, рекомендуется предварительно изучить нечеткие модели S-, D- и T- модели в разделах 3.3, 3.3.3-3.3.5, 5.7 в книге: Афанасьева Т. В., Ярушкина Н. Г. Нечеткое моделирование временных рядов и анализ нечетких тенденций, 2009.

D- и T-модели являются модификациями S-модели и, также, как и S- модель, основаны на генерации по временному ряду нечетких продукционных правил, которые затем используются для прогнозирования будущих значений.

Моделирование нечетких временных рядов в соответствии с нечеткой моделью, предложенной в работе Song, 1993a, состоит в реализации следующих шагов:

1. Определение нечетких переменных – разбиение диапазона данных временного ряда на множество интервалов (носителей нечетких множеств), определение для каждого диапазона лингвистических значений нечетких множеств и их функций принадлежности.
2. Формирование логических отношений
3. При этом зависимость в нечетких значениях может быть не только в виде зависимости текущего значения от предыдущего, но и от p-го предыдущего значения (значение p называют порядком нечеткой модели): $Yt = (Yt\text{-}1 \times Yt\text{-}2 \times \ldots \times Yt - p) \circ R(t, t - p)$
4. Фаззификация входных данных – определение степени принадлежности входных данных входным нечетким переменным.Вычисление результата применения нечеткого правила $Rij\ (t, t\text{-}1)$ для каждой импликации
5. Вычисление результирующего отношения R как объединением $\bigcup_{i,j} Rij\ (t, t\text{-}1)$.
6. Применение полученной модели к входным данным и получение выходных нечетких результатов.
7. Дефаззификация нечетких результатов, например, вычисляя центр тяжести.

Обычно для вычисления приближенного численного решения в этом случае применяют алгоритм Мамдани, также, как и общий подход к прогнозированию временных рядов на основе нечетких моделей (раздел 3), книги Ярушкина Н. Г., Афанасьева Т. В., Перфильева И. Г. Интеллектуальный анализ временных рядов: учебное пособие – М.: ИД «ФОРУМ» ИНФРА-М, 2012.

Рекомендуемая последовательность выполнения работы

При помощи системы salx.pw найти оптимальную нечеткую модель прогнозирования из представленных на веб-сайте при помощи следующего алгоритма:

1. Выбрать временной ряд (выбирать из full – *, так как на втором шаге в лабораторной работе № 5 было выполнено объединение данных с веб-сайта CIF_2016);
2. Указать глубину прогноза q как горизонт прогнозирования для каждого ВР (обычно q составляет 10% длины временного ряда);
3. Выбрать одну из нечетких моделей: S-, D- или T-модель;
4. Выбрать порядок p нечеткой модели: p = 1, 2,…;
5. Для каждого порядка модели выбрать тестовый отрезок временного ряда: либо глубину прогноза q, либо удвоенную глубину прогноза 2*q;
6. Выбрать характеристики нечеткой модели, для которых внешний критерий качества SMAPE минимален, для этой модели и для выбранного порядка провести прогнозирование вперед на q значений и оценить визуально качество модели. Если характер прогнозного временного ряда соответствует характеру исходного ряда, то зафиксировать эту модель как лучшую;
7. Повторить для каждой нечеткой модели;
8. Выбрать лучшую нечеткую модель с характеристиками для конкретного временного ряда с минимальным внешним SMAPE и визуальным сходством;
9. Зарегистрировать результаты в таблице результатов лабораторной работы.

6.3. Задание к лабораторной работе

1. Познакомиться с возможностями прогнозирования ВР с помощью системы: http://salx.pw/IFSA
2. Провести прогнозирование и выбрать лучшую нечеткую модель заданных ВР, изменяя параметры (порядок модели, тестовый отрезок) в системе http://salx.pw/IFSA по критерию минимума Внешний SMAPE.

Результаты занести в таблицу.

6.4. Контрольные вопросы

При защите лабораторной работы необходимо ответить на три вопроса из списка контрольных вопросов:

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.

Лабораторная работа №7. Проведение сравнительного анализа моделей временных рядов

7.1. Цель работы и общие требования

Цель работы: научиться проводить сравнительный анализ по точности прогнозирования и визуальному соответствию для выбора адекватной модели прогнозирования временного ряда.

7.2. Методические рекомендации и материалы

Для проведения сравнительного анализа необходимо использовать результаты, полученные при выполнении лабораторных работ № 5 и 6.

Исходные данные: заполненная таблица результатов.

Результаты: текстовый отчет о прогнозировании временных рядов по соответствующему варианту статистическими и нечеткими моделями с выводами и рекомендациями, какую модель целесообразно использовать и почему. Привести критерии точности и показать на графиках визуальное соответствие рекомендованных моделей.

Ход работы:

1. Для каждого временного ряда заданного варианта заполнить отчет снимками экрана с приведением характеристик рекомендуемой модели с оптимальными характеристиками.
2. Привести итоговую таблицу, в которой расположить рекомендуемые модели.
3. Оформить текстовую часть отчета в соответствии с требованиями.
4. Отправить текстовую часть отчета и таблицу результатов лабораторной работы на проверку.

7.3. Задание к лабораторной работе

Провести сравнительный анализ данных полученных результатов из лабораторных работ № 5 и 6 на основе свойств ВР.

7.4. Контрольные вопросы

При защите лабораторной работы необходимо ответить на три вопроса из списка контрольных вопросов:

1. Постановка задачи, основные задачи анализа ВР. Критерии качества моделей.
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).
6. Модели случайной компоненты ВР (AR, MA, ARMA, ARIMA).
7. Модели сезонных колебаний (индексные методы, адаптивные методы EST, спектральные методы, сезонная Arima).

8. Нечеткий подход к прогнозированию ВР. Этапы анализа и прогнозирования.

9. Методы прогнозирования ВР в нечетком подходе. Базовая модель нечеткого ВР Q. Song & B. Chissom (S-модель) и ее разновидности.

10. Виды моделей нечеткого логического вывода, применяемые при прогнозировании нечетких ВР (Мамдани, Суджено).

11. Задача анализа нечетких тенденций ВР. Формализация нечеткой тенденции. Виды нечетких тенденций.

12. Основные задачи анализа ВР в терминах нечетких тенденций.

13. Возможности перехода к лингвистическим ВР в нечетком подходе.

14. Проблемы и преимущества прогнозирования ВР в нечетком подходе.

15. Примеры задач прогнозирования в решении прикладных задач.

Лабораторная работа №8. Прогнозирование временных рядов на языке R

8.1. Цель работы и общие требования

Цель работы. Научиться использовать готовые решения по прогнозированию временных рядов на языке R.

8.2. Методические рекомендации и материалы

Исходные данные:

Материалы сайта об интеллектуальном анализе данных: http://www.rdatamining.com/

Сайт с документацией о языке R: https://www.rdocumentation.org

Результат: текстовый отчет, содержащий информацию о ходе выполнения работы и о результатах прогнозирования с выводом критериев точности.

Ход выполнения работы

Для выполнения работы необходима установка среды вычисления R версии 3.4.3 и IDE RStudio версии 1.1.383. Дальнейшая работа рекомендована в IDE.

Первым шагом установите пакеты с помощью следующих команд:

1) ***install.packages("zoo")*** # нерегулярные временные ряды
2) ***install.packages("forecast")*** # ARMA, экспоненциальное сглаживание

После установки всех необходимых пакетов создайте скрипт:

- Загрузить данные из csv-файла (был выбран ряд ts25, использованный в лабораторных 1 и 2)
- Преобразовать данные во временной ряд
- Провести прогноз временного ряда
- Вывести график с прогнозом
- Рассчитать оценки качества прогноза

Листинг скрипта приведен ниже:

```
require(forecast) require(zoo)
tsFile <- read.zoo(file = "D:/Projects/R/IADiP/ts25.csv", sep = ";", header = TRUE, tz = "", format = "%d.%m.%Y", index.column = "Date") tsData <- ts(data = tsFile)
fit <- auto.arima(tsData) fcast <- forecast(fit) plot(fcast) accuracy(fcast)
```

Для загрузки данных из файла используется пакет zoo. Он автоматически преобразует даты, извлеченные из документа, в экземпляры класса Date. Структура полученного объекта представлена на рисунке 8.1.

```
> tsFile
1980-01-01 1980-01-02 1980-01-03 1980-01-04 1980-01-05 1980-01-06 1980-01-07 1980-01-08
  699.3228   773.0060   653.9413   717.4517   799.6718   696.5065   846.1512   622.0792
1980-01-09 1980-01-10 1980-01-11 1980-01-12 1980-01-13 1980-01-14 1980-01-15 1980-01-16
  642.2988   618.4263   774.0116   845.6055   773.8636   680.8580   827.1419   830.9161
1980-01-17 1980-01-18 1980-01-19 1980-01-20 1980-01-21 1980-01-22 1980-01-23 1980-01-24
  876.7645   543.8172   693.0192   812.1436   839.9288   833.1079   745.2970   950.1273
1980-01-25 1980-01-26 1980-01-27 1980-01-28 1980-01-29 1980-01-30 1980-01-31 1980-02-01
  647.0077   818.3905   805.7073   638.6870   719.7258   838.9916   746.5586  1028.3295
```

Рис. 8.1. Структура объекта tsFile

Далее, извлеченные данные преобразуются в объект класса ts.

Структура полученного объекта представлена на рисунке 8.2.

```
Time Series:
Start = 1
End = 120
Frequency = 1
  [1]  699.3228  773.0060  653.9413  717.4517  799.6718  696.5065  846.1512  622.0792
  [9]  642.2988  618.4263  774.0116  845.6055  773.8636  680.8580  827.1419  830.9161
 [17]  876.7645  543.8172  693.0192  812.1436  839.9288  833.1079  745.2970  950.1273
 [25]  647.0077  818.3905  805.7073  638.6870  719.7258  838.9916  746.5586 1028.3295
 [33]  641.5209  756.6558  953.0441  733.1484  899.1895  759.7470  847.3639  859.3450
 [41]  968.9986  864.8138  864.2475  767.1944  607.9178  737.0656  618.1021  757.1793
 [49]  920.5465  797.8954  812.7431  909.1448  993.8944  799.5641  958.6595 1006.8544
 [57]  944.0317 1018.0137  999.7774  848.7164  982.0268  841.2505  938.6457  896.9882
 [65]  988.9585 1000.0165 1065.6241  976.0958 1016.7573  929.3219  885.9090 1124.6346
 [73] 1054.3197  937.9377  977.1270  885.7179 1083.6657  920.9896 1195.6729  993.6415
 [81]  979.2849  985.9075 1005.7260  945.5448 1114.4066 1124.5896 1050.8174  998.9414
 [89] 1130.7146 1013.2595 1115.5934 1181.2886 1074.1177 1075.2162  986.7974 1109.7244
 [97] 1205.5753 1063.4691 1165.6841 1132.1437 1090.4403 1195.4315 1092.3591 1077.6837
[105] 1199.2171 1169.0670 1194.5785 1135.9090 1152.4119 1280.9281 1122.8497 1244.0913
[113] 1092.2180 1114.3269 1099.4740 1214.8587 1093.2109 1103.6788 1313.0685 1383.1386
attr(,"index")
  [1] "1980-01-01 MSK" "1980-01-02 MSK" "1980-01-03 MSK" "1980-01-04 MSK" "1980-01-05 MSK"
  [6] "1980-01-06 MSK" "1980-01-07 MSK" "1980-01-08 MSK" "1980-01-09 MSK" "1980-01-10 MSK"
 [11] "1980-01-11 MSK" "1980-01-12 MSK" "1980-01-13 MSK" "1980-01-14 MSK" "1980-01-15 MSK"
 [16] "1980-01-16 MSK" "1980-01-17 MSK" "1980-01-18 MSK" "1980-01-19 MSK" "1980-01-20 MSK"
 [21] "1980-01-21 MSK" "1980-01-22 MSK" "1980-01-23 MSK" "1980-01-24 MSK" "1980-01-25 MSK"
 [26] "1980-01-26 MSK" "1980-01-27 MSK" "1980-01-28 MSK" "1980-01-29 MSK" "1980-01-30 MSK"
 [31] "1980-01-31 MSK" "1980-02-01 MSK" "1980-02-02 MSK" "1980-02-03 MSK" "1980-02-04 MSK"
 [36] "1980-02-05 MSK" "1980-02-06 MSK" "1980-02-07 MSK" "1980-02-08 MSK" "1980-02-09 MSK"
```

Рис. 8.2. Структура объекта tsData

Для прогнозирования может быть использован любой метод, например, на основе модели ARIMA (подробный список параметров можно найти по ссылке), хотя так же можно было использовать и другие модели (функции имеют аналогичный синтаксис), например, TBATS. Параметры (p, d, q) и (p1, d1, q1) подбираются автоматически с помощью функции auto.arima() на основе критериев AIC и BIC (в случае отсутствия сезонности, (p1, d1, q1) устанавливается нулями: (0, 0, 0)). Так же их можно установить вручную, используя функцию Arima(). Для данного временного ряда была автоматически выбрана модель ARIMA(0,1,1).

Результаты прогноза представлены на рисунке 8.3, где по оси X отмечены порядковые номера данных, а по оси Y – значения.

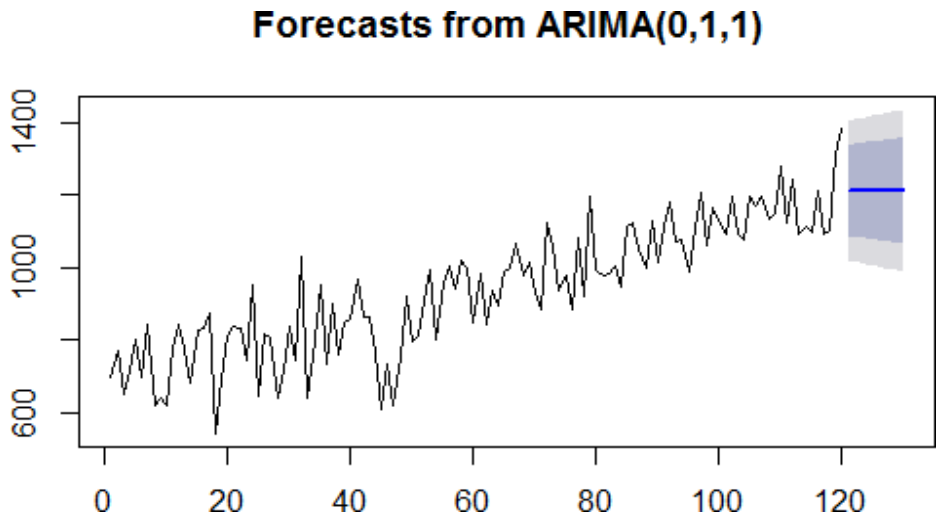

Рис. 8.3. Прогноз временного ряда

Для определения точности прогноза используется функция accuracy3.

Результаты ее работы представлены на рисунке 8.4.

```
> accuracy(fcast)
                    ME      RMSE       MAE       MPE      MAPE      MASE        ACF1
Training set 20.99909 96.92638 78.83672 1.083851 8.848642 0.7585133 -0.0979381
```

Рис. 8.4. Критерии качества прогноза

8.1. Задание к лабораторной работе

С помощью языка R спрогнозировать произвольный временной ряд и получить оценки качества прогноза.

8.2. Контрольные вопросы

- Назначение пакета R.
- Какие модели временных рядов реализованы в R?
- Что необходимо выполнить, чтобы провести прогнозирование в R?

Какие критерии точности можно получить в R? Что они характеризую

СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ

1. Халеева Е.П. Анализ данных средствами языка R : учебное пособие / Халеева Е.П., Аль-Ханани М.А., Лютикова М.Н.. — Саратов : Вузовское образование, 2022. — 71 c.
2. Структуры и алгоритмы компьютерной обработки данных : учебно-методическое пособие для проведения лабораторных работ / Ю.М. Мартынюк [и др.].. — Тула : Тульский государственный педагогический университет имени Л.Н. Толстого, 2021. — 73 c.
3. Калачева Н. М. Информационные технологии [Текст]: учебник для студентов высших учебных заведений / Н. М. Калачева, С. С. Кравчук. – 3-е изд., перераб. и доп. – Москва: Издательство «Юрайт», 2019. – 280 с.
4. Коршунов М. К. Экономика и управление: применение информационных технологий: учебное пособие для вузов / М. К. Коршунов; под научной редакцией Э. П. Макарова. — 2-е изд. — Москва : Издательство «Юрайт», 2022. — 110 с.