ДИСЦИПЛИНА	**Проектирование интеллектуальных систем (часть 1/2)**
ИНСТИТУТ	**информационных технологий**
КАФЕДРА	**вычислительной техники**
ВИД УЧЕБНОГО МАТЕРИАЛА	**Материалы для практических/семинарских занятий**
ПРЕПОДАВАТЕЛЬ	**Холмогоров Владислав Владиславович**
СЕМЕСТР	**6, 2023-2024**

Практическая работа № 7

«Интеллектуальный анализ данных»

по дисциплине «Проектирование интеллектуальных систем (часть 1/2)»

Цели: приобрести навыки проведения полного цикла интеллектуального анализа данных.

Задачи:

1) Сделать полный цикл интеллектуального анализа данных – от сбора данных и их предобработки, до обучения моделей и создания полноценного приложения, выполнив следующие пункты:

– определить предметную область решаемой задачи, может совпадать с таковыми из предыдущих работ в соответствии с тематикой заданий;

– найти или сгенерировать набор данных, содержащий данные для задачи выбранной предметной области, соблюдая особенности сбора данных (data collection), в соответствии с чем описать (по возможности) особенности набора данных, достоверность, способы сбора, обеспечения и контроля качества данных в наборе (см. Примечание 1);

– выполнить предобработку данных в соответствии с определённым качеством данных (см. Примечание 2);

– в соответствии с задачей предметной области реализовать и обучить (при необходимости) модель/модели, либо повторно обучить модель из предыдущих практических работ, сравнить результаты работы модели, обученной на непредобработанных и предобработанных данных, а также с результатами предыдущей работы (если задача и модель аналогичные) с помощью соответствующих метрик качества и графиков;

В качестве дополнительного задания реализовать хотя бы один из следующих пунктов (один – на оценку 4, все – на оценку 5):

– выполнить аугментацию данных (см. Примечание 3);

– реализовать приложение для взаимодействия с моделью, включающее программные и пользовательский интерфейсы (см. Примечание 4);

ПРИМЕЧАНИЕ:

1) Data collection (сбор данных) является одним из основных этапов в IT-отрасли, так как соблюдение его теории позволяет собирать и находить качественные наборы данных, которые в дальнейшем требуют меньше действий, связанных с их предобработкой, и оказывают меньшее негативное воздействие на инструменты их обработки и результаты этой обработки. В теории сбора данных выделяют следующие аспекты:

– сбор первичных и вторичных данных: первый способ предполагает сбор исходных данных непосредственно из источника или посредством прямого взаимодействия с респондентами, к нему относятся интервью, опросы и анкетирование, наблюдения, эксперименты, фокус-группы, второй предполагает использование существующих данных, собранных кем-то другим, для целей, отличных от первоначального намерения, к нему относятся опубликованные источники, онлайн базы данных, правительственные и институциональные (то есть официальные источники) документы, общедоступные данные и результаты прошлых исследований;

– инструменты сбора данных: опросы (личные, онлайн, ассоциативные и телефонные), ролевые игры, наблюдения, датчики и телеметрия;

– типичные проблемы при сборе данных: противоречия в данных, время простоя данных, неоднозначность данных, дублирующие и большие данные, неточные и скрытые данные, релевантные данные;

– обеспечение и контроль качества (проблема целостности данных): обеспечение качества (quality assurance) данных представляет собой комплекс действий, выполняемый перед сбором данных и направленный на создание универсального протокола сбора данных (как единой профилактики), всестороннего и подробного описания процедур сбора данных (например, чёткое определённые периоды и сроки собираемых данных, полный список собираемых объектов, подробное описание инструментов сбора данных, чёткие инструкции по использованию и механизмы документирования), контроль качества (quality control) представляет собой комплекс действий,

направленных на наблюдение и исправление ошибок сбора данных, которые заключаются в документировании процессов сбора, мониторинге, действиях по исправлению ошибок при сборе данных и минимизации будущих ошибок (к таковым могут относится погрешности наблюдений, нарушение протоколов, мошенничество и ошибки отдельных элементов данных).

2) Предобработка (предварительная обработка) данных представляет собой комплекс действий по преобразованию набора данных с целью обеспечения качества и возможности их дальнейшего анализа (самой обработки), включает в себя следующие методы:

– очистка данных (Data Cleaning): выявление и исправление ошибок или несоответствий в данных, включает в себя обработку отсутствующих значений (удаление пропусков, дубликатов и оценка недостающих значений), обработку зашумлённых данных (кластеризация, интерполяция, биннинг, кэппинг, тримминг и т.д.);

– интеграция данных (Data Integration): объединение данных из нескольких источников для создания единого набора данных, выполняется с помощью слияния данных, сопоставление схем данных и связывания однородных записей;

– преобразование данных (Data Transformatio): преобразование данных в подходящий для анализа формат, обычно выполняется с помощью нормализации, стандартизации и дискретизации (с помощью биннинга по ширине и частоте и кластеризации);

– сокращение данных (Data Reduction): предполагает уменьшение размера набора данных при сохранении важной информации, выполняется с помощью выбора признаков (Feature Selection, к нему относятся корреляционный и частотный анализы, а также анализ главных компонент), извлечение признаков (Feature Extraction, к нему относятся факторный анализ, латентное размещение Дирихле и кластерный анализ), сэмплинг (на основе случайной, стратифицированной или систематической выборки) и сжатие данных (сжатие изображений, вейвлет-сжатие или сжатие в архивы данных).

3) Аугментация данных – метод искусственного увеличения обучающего набора путем создания модифицированных копий набора данных с использованием существующих данных, выполняется для увеличения точности обучаемой модели и предотвращения переобучения. Так как табличные данные отличаются от обычных текстовых данных, и тем более от фото и аудио данных, им не подойдут классические варианты вроде перестановки слов, геометрических преобразований и спектральных сдвигов, для них используются следующие отдельные методы:

– генеративные модели нейронных сетей: генеративно-состязательные сети и автоэнкодеры;

– модельные генеративные методы: SMOTE, Borderline SMOTE, ADASYN, MixUp, TabDDPM;

– функции генерации типичных, статистических или закономерных значений (при определённой достаточной точности аппроксимации данных или известного факта о первоначальной функциональной зависимости).

4) В реальных проектах интеллектуальные решения с учётом популярности микросервисной архитектуры и возрастающим размером разрабатываемых систем обычно являются лишь одним (хоть иногда и главным) компонентом системы (например, те же модули принятия решений или элементы бизнес-интеллекта), что нужно учитывать при их создании, поэтому для обеспечения возможности интероперабельности (функциональной совместимости) необходимо создавать интерфейсы для программного взаимодействия (то есть создавать API) с другими программными компонентами, вне зависимости от их функциональности, языка и способа передачи данных, а также создавать пользовательские интерфейсы для обеспечения возможности прямого управления интеллектуальными модулями и возможности реализации определённых видов тестирования (например, юзабилити тестирования, тестирования вариантов использований или автоматизированного тестирования ботами).

ОСОБЫЙ БОНУС (доступен только в том случае, если выполнены пункты 2-го задания):

Создать единое приложение с пользовательским и программным интерфейсами для выполненных (это значит, что необязательно для всех, но не менее трёх главных, названных дальше) практических работ курса предмета с возможностью загрузки и ввода набора данных, его предобработкой и анализом на основе хотя бы кластеризации, классификации и регрессии (это минимальное, но и достаточное условие для выполнения данного пункта).