Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Применение возможностей языка Python в алгоритмах кластеризации и прогнозирования временных рядов

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

по направлению 09.04.01 «Информатика и вычислительная техника»
студента 2 курса 271 группы
факультета компьютерных наук и информационных технологий
Султанова Джамала Артуровича

Научный руководитель
к.ф.-м.н., доцент _____ В.А. Поздняков

Заведующий кафедрой
к.ф.-м.н., доцент _____ Л.Б. Тяпаев

Саратов 2018

Введение. Развитие технологий привело к увеличению количества накапливаемой и обрабатываемой машинами информации. Устройства с доступом к сети Интернет становятся всё дешевле и доступнее, а крупные игроки на рынке информационных технологий внедряют сеть во все сферы человеческой жизни.

Рост количества генерируемой информации привел к появлению нескольких направлений в сфере изучения данных, среди которых Data Mining – исследование данных на предмет нетривиальных знаний, которые могут быть полезны человеку.

Для написания данной работы были изучены средства анализа данных в языке Python. Полученные знания были использованы для анализа данных из образовательной системы Moodle, а также для прогноза количества публикаций на сетевом ресурсе по открытым данным о количестве публикаций за весь период существования ресурса.

Объект исследования: реализация методов интеллектуального анализа данных средствами языка Python.

Цель работы – сравнительный анализ внедренных методов интеллектуального анализа данных с методами, предлагаемыми языком Python.

Задачи, поставленные на период исследования:

- знакомство с задачами интеллектуального анализа данных и методами их решения;
- изучение и подготовка данных для проведения интеллектуального анализа;
- изучение возможностей Python, используемых в анализе данных;
- реализация моделей интеллектуального анализа на языке Python и сравнительный анализ с актуальными моделями.

За последние несколько лет работа с данными для поиска в них скрытых знаний стала одной из самых актуальных задач в индустрии. Data Mining внедряется во многие сферы, например в научные исследования, медицину и

здравоохранение, в трейдинг и финансовую сферу, и прочие. Анализ данных используется в сфере образования и предоставляет возможность на основе отчетов из систем управления обучением создавать улучшающие процесс получения знаний рекомендации для студентов и преподавателей.

Помимо сферы образования, средства анализа данных находят практическое применение в прогнозировании временных рядов.

Работа состоит из трех глав:

1. В первой главе рассмотрены основные понятия интеллектуального анализа данных, задачи классификации, кластеризации и прогнозирования, а так же методы решения этих задач.
2. Во второй главе рассмотрены средства языка Python, использующиеся в решении задач интеллектуального анализа данных.
3. Третья глава содержит решение поставленных задач с применением языка Python.

Постановка задачи и изучение теории. Data Mining — совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

В данной работе необходимо применить методы анализа данных для оценки сложности вопросов, определить похожие по сложности вопросы, а также распределить студентов по уровню знаний. Эти данные могут быть использованы как база для создания рекомендаций.

Помимо сферы образования, средства анализа данных находят практическое применение в прогнозировании временных рядов. На основе такого прогноза можно определить количество серверов, которые могут понадобиться сетевым сервисам в будущем, предположить спрос на товары или котировки курсов. Для предсказания было разработано несколько подходов: нейронные сети, регрессионные модели, ARIMA и т.д. В данной работе будет произведен сравнительный анализ используемых методов с

реализованной на языке Python open-source библиотекой Prophet.

К этапам построения модели интеллектуального анализа данных относят[1]:

1. Постановка задачи

Этап включает в себя накопление требований, определение проблемной области и цели анализа.

2. Подготовка данных

На этом этапе происходит обращение к источнику данных и их последующее получение.

3. Изучение данных

Необходимо выяснить, описывает ли исходная информация рассматриваемую предметную область. Так же изучая данные, можно обнаружить ошибки в них. К изучению данных относят: расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и пр. Такие методы позволяют принять решение о готовности к построению модели. Рассчитав величину стандартного отклонения мы можем определить точность результатов. Если эта величина превышает норму, могут потребоваться новые данные.

4. Построение модели

На этом этапе создается структура ИАД, которая в дальнейшем будет обрабатываться. Под процессом обработки подразумевается выбор и использование алгоритма Data Mining на наборе данных для выявления искомых закономерностей.

5. Исследование и проверка моделей.

Существует два способа проверки моделей – тестирование на специальном наборе данных, который подходит для предсказательных задач и перекрестная проверка, суть которого состоит в создании нескольких подмножеств и запуске модели на каждом подмножестве для сравнения результатов.

Задача кластеризации. Кластеризация используется для анализа некоторого набора данных. Её задачей является разбиение множества объектов на группы, имеющие общие свойства или характеристики. Такие группы называются кластерами. Решением задачи является распределение объектов по кластерам. Методы кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

К иерархическим методам относят такие алгоритмы интеллектуального анализа, которые в результате своей работы строят полное дерево вложенных кластеров (дендрограмма). Дендограмма – диаграмма, наглядно представляющая близость кластеров друг к другу.

Иерархические методы кластерного анализа бывают двух типов:

- агломеративные (объединительные). Выполняют поочередное объединение исходных элементов. То есть изначально каждый элемент представляет отдельный кластер, которые затем объединяются, уменьшая количество кластеров (построение снизу вверх);
- дивизимные (или разделяющие). Обратные агломеративным. Изначально все элементы содержатся в одном кластере, а затем исходный кластер разделяется и образуется последовательность групп[2].

Неиерархические методы отличаются тем, что количество кластеров определяется изначально. В ходе неиерархической кластеризации набор данных разделяется по выбранным кластерам так, чтобы целевая функция алгоритма достигала минимума. Одним из самых популярных методов неиерархической кластеризации стал алгоритм k-средних.

Средства интеллектуального анализа данных в Python. Кластеризация используется для анализа некоторого набора данных. Её задачей является разбиение множества объектов на группы, имеющие общие свойства или характеристики. Такие группы называются кластерами. Решением задачи является распределение объектов по кластерам. Методы кластерного

анализа можно разделить на две группы:

Scikit-learn – библиотека, реализующая алгоритмы ИАД. В то время как Pandas и NumPy служат для загрузки данных и действий над ними, Scikit-learn дает разработчику инструменты для моделирования данных. Библиотека предоставляет следующий функционал:

- НБК, нейросети, метод моделирования на опорных векторах, дерево принятия решений и др;
- Кластеризация;
- Перекрестная проверка (используется для оценки эффективности работы модели);
- Большое количество встроенных наборов данных, которые можно использовать для тестирования.

Одной из задач текущей работы является кластеризация. То есть разбиение данных на кластеры – группы, объединяющие элементы со схожими признаками. В библиотеке Scikit-learn реализовано множество алгоритмов кластеризации. Рассмотрим принцип работы алгоритма k-средних:

1. Выбирается некоторое количество кластеров;
2. Определяются центроиды;
3. Для каждого элемента исходного набора вычисляется расстояние до центроидов;
4. Выбирается кратчайшее расстояние;
5. Центроид пересчитывается так, чтобы он стал центром подмножества элементов, отнесенных к этому центроиду;
6. Шаги 3-5 повторяются заданное фиксированное количество раз

Инструменты для прогнозирования. Помимо задачи кластеризации, в рамках данной работы рассматривается возможность прогнозирования временных рядов при помощи методов интеллектуального анализа данных. Для решения этой задачи была выбрана библиотека Prophet, которая не только

строит прогнозы, но и позволяет улучшать прогноз, редактируя простые параметры, то есть облегчая работу аналитиков.

Сравнительный анализ кластеризации средствами Python и SQL Server. По результатам исследования, большинство объектов в кластерах, полученных при помощи SQL Server совпадают с объектами кластеров из Python. Различие объясняется разными алгоритмами кластеризации – в Python использовался K-means, а в SQL Server используется метод максимизации ожиданий (EM-алгоритм). При этом определено, что точность моделей созданных на Python выше. Точность определяется при помощи перекрестной проверки.

Сравнительный анализ кластеризации средствами Python и SQL Server. Рассмотрев две модели прогнозирования – устоявшуюся ARIMA и принципиальной новую и практически не используемую Prophet, можно сделать вывод, что при приблизительно одинаковой точности, Prophet требует в разы меньше усилий и времени на прогноз. При этом в библиотеку заранее включены данные, способные влиять на прогноз, например праздничные дни, и их использование может существенно повлиять на прогнозы. К недостаткам библиотеки относится малое количество информации о принципах работы и практически полное отсутствие документации.

Заключение. Язык Python предоставляет широкий выбор возможностей для проведения аналитики над данными, полученными из различных сетей. В рамках исследования этих возможностей, были изучены задачи ИАД, методы их решения и средства языка, предоставляющие эти возможности.

Использование интеллектуального анализа данных помогает улучшить качество предоставления различных видов услуг. Методы кластерного анализа помогают сегментировать студентов на группы по общему уровню знаний, а не только по итоговой оценке; при помощи методов классификации можно проводить подробный анализ текстов и автоматически определять наличие спама или рекламы; методы прогнозирования временных рядов способствуют

улучшению серверной части сетевых приложений, и др.

Решение задач ИАД средствами языка Python позволяет уделять меньше времени подготовке исходных данных (Pandas) и работе с ними (NumPy), а также облегчает сам процесс анализа (Scikit-learn, Prophet).

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Нестеров С.А. Базы данных. Интеллектуальный анализ данных. Учебное пособие. / С.А. Нестеров. - СПБ.: Изд-во Политехн. ун-та, 2011. – 272 с.

2 Чубукова И.А. Data Mining. Задачи Data Mining. Классификация и кластеризация [Электронный ресурс] URL: http://www.intuit.ru/studies/courses/6/6/lecture/166?page=4 (дата обращения 17.05.2018).