

Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра технологий
программирования

**СИСТЕМА ИНФОРМАЦИОННОГО АНАЛИЗА ДЕТАЛИЗАЦИИ
ТЕЛЕФОННЫХ СОЕДИНЕНИЙ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

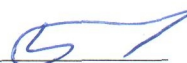
студентки 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные
технологии
факультета КНиИТ
Жадаевой Анастасии Петровны

Научный руководитель
к. ф.-м. н.


10.6.17

И. А. Батраева

Заведующий кафедрой
к. ф.-м. н.


10.6.17

И. А. Батраева

Саратов 2017

ВВЕДЕНИЕ

В связи с продолжающимся ростом населения и прочими социальными факторами, люди по всему миру становятся менее зависимыми от проводного телефона, а количество пользователей и объем трафика мобильной стремительно увеличивается.

Согласно данным, опубликованным аналитическим агентством AC&M Consulting, количество абонентов услуг сотовой связи (по числу SIM-карт) в России по итогам II квартала 2016 г. увеличилось до 251,6 млн.

Такой колоссальный объем данных требует создания специального программного обеспечения по его обработке и анализу.

Особый интерес в области исследований представляет классификация информации по различным критериям.

Для таких сфер, как маркетинг и обеспечение безопасности, первостепенным является решение такой нетривиальной задачи, как выявление общих связей между несколькими устройствами и близкого круга общения пользователя.

Целями данной работы являются разработка программного комплекса импорта данных, а также реализация методов, позволяющих получить анализ биллингов, в частности частотные характеристики операций клиентов и меж-клиентские связи.

Дипломная работа состоит из введения, 2 разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы — 73 страницы, из них 42 страницы — основное содержание, включая 20 рисунков и список использованных источников из 24 наименований.

1 Краткое содержание

Во введении выдвигаются аргументы в подтверждение актуальности темы и ставятся цели: разработка программного комплекса импорта данных, а также реализация методов, позволяющих получить анализ биллингов, в частности частотные характеристики операций клиентов и межклиентские связи.

В первом подразделе первой главы «Предметная область» представлены общие сведения о рассматриваемой предметной области. Исследуемые в данной работе методы проведения частотного анализа и выявления закономерностей относятся к области аналитики, которая в свою очередь тесно связана с такими понятиями как Data Mining, статистика, машинное обучение и др.

Data Mining — процесс поиска новых, полезных в практической деятельности и доступных для интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Благодаря развитию методов обработки и хранения данных, технология стала усовершенствованной версией интеллектуального анализа.

Согласно [1] Data Mining — это научная область, которая во многом основывается на понятиях и методах теории баз данных, прикладной статистики, распознавания образов, алгоритмизации и др.

В ходе решения поставленных в работе задач, возникла необходимость сбора выборочных данных об активности клиентов. Для этого были применены методы подготовки данных, присущие статистике.

Статистика — это наука, разрабатывающая набор приемов и способов сбора, обработки и анализа информации. Статистика представляет собой систему из методов планирования эксперимента, сбора, анализа и получения выводов на основании этих данных.

Кроме определения этих дисциплин, рассмотрены также их роль в решении поставленной задачи выявления общих связей между несколькими устройствами и близкого круга общения пользователя.

Второй подраздел первой главы носит название «Методы машинного обучения», в нем приведены общие определения из области машинного обучения, которые используются в дипломной работе. В частности рассмотрены типы и классификация алгоритмов машинного обучения.

Машинное обучение — это изучение методов, основанных на данных, способных имитировать, понимать и помогать решать задачи обработки ин-

формации. В этом стремлении возникают многие связанные с этим вопросы, такие как сжатие данных, их интерпретация и обработка.

Зачастую эти методы направлены не на подражание действиям человека, а скорее на их улучшение, например, для наиболее быстрого получения информации.

В [2] говорится, что эта дисциплина базируется на математических принципах, использует статистические и оптимизирующие методы. Её отличительной чертой является решение задач, вычисляющих эффективность и обеспечивающих переобучение. Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам.

Особый интерес для данной работы представляет тип машинного обучения с учителем. Обучение с учителем базируется на разработке функции, которая отображает входные данные для заданных выходных. Одной из стандартных формулировок контролируемой задачи обучения является проблема классификации: учитель должен обучить функцию, чтобы она смогла отображать вектор в один из нескольких классов на основе поиска по данным примерам ввода-вывода.

Для решения поставленной в дипломной работе задачи выявления общих связей между несколькими устройствами и близкого круга общения пользователя, и достижения наиболее качественного результата оптимальным решением стало применение алгоритма классификации.

Классификация — один из разделов машинного обучения, посвященный решению следующей задачи распределения объектов по классам. Обычно алгоритм, основываясь на обучающей выборке, т. е. множестве объектов с предопределенными классами, должен определить классовую принадлежность остальных объектов.

В качестве выходных данных в таком случае будет выступать название или идентификатор искомого класса для некоторого подмножества объектов.

В математической статистике задачи классификации называются также задачами дискриминантного анализа.

В работе используется метод *k*-ближайших соседей. Его основная идея заключается в определении класса объекта на основе примера из обучающей выборки. Несмотря на простоту этой идеи, данный метод показывает хорошие результаты работы и широко применяется в различных областях:

1. Приложения компьютерного зрения, в том числе распознавание символов и лиц, как в неподвижных изображениях, так и в видео.
2. Прогнозирования релевантности поисковых запросов.
3. Выявления закономерностей в генетических данных, возможно, их использования при выявлении специфических заболеваний.
4. Обнаружения афер. Новые случаи мошенничества могут быть сходны с уже произошедшими попытками обмануть системы безопасности. Алгоритм k-ближайших соседей может распознать их для дальнейшего рассмотрения.
5. Медицины. Согласно [3], рассматриваемый метод может классифицировать пациентов по различным показателям, основываясь на уже приобретенном опыте.

В целом, классификатор ближайших соседей хорошо подходит для задач классификации, где отношения между объектами и целевыми классами многочисленны, сложны или трудны для понимания, но элементы подобного типа классов имеют тенденцию быть довольно однородными.

К плюсам данного метода можно отнести следующее:

- Простота и эффективность.
- Данный метод не делает предположений о базовом распределении данных.
- Быстрый этап обучения.

В книге «Machine Learning with R: Second Edition» [4] авторы выделяют следующие минусы:

- Требуется выбора подходящего параметра k .
- Недостающие данные требуют дополнительной обработки.
- Медленный этап классификации.

Метод k-ближайших соседей — один из методов решения задачи классификации. Предполагается, что на начальном этапе уже имеется некоторое множество объектов с заданной классификацией. На основе этих данных вырабатывается правило, позволяющее отнести новый объект к одному из возможных классов: объект считается принадлежащим тому классу, к которому относится большинство его ближайших соседей. Под «соседями» здесь понимаются объекты, близкие к исследуемому.

Под близостью в данном случае подразумевается сходные характери-

ки, например, частота звонков в день, число смс определенному человеку и прочее.

В качестве расстояния между объектами выступает Евклидова норма, рассчитываемая по формуле $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$, где (x_1, y_1) и (x_2, y_2) — координаты, определенные на основе значений численных параметров объектов.

Отдельной трудоемкой задачей является выбор значения параметра k , от которого напрямую зависит результат применения данного метода классификации.

Проблема выбора k заключается в том, что при малом значении, его ближайшим соседом может оказаться объект с неверно определенным классом, слишком большое значение также приведет к некорректной классификации.

Таким образом, крайние значения k не приведут к получению верного результата классификации. Согласно [5], на практике оптимальное значение параметра k определяют по критерию скользящего контроля, чаще всего — методом исключения объектов по одному (leave-one-out cross-validation).

Однако в общем случае, k выбирают равным $\sqrt{N}$, где N — это число объектов, рассматриваемых в качестве соседей. Однако при уверенности в «чистоте» выборки, k может принимать меньшее значение. Кроме того, для решения этой проблемы более близким объектам приписывается больший вес.

Подраздел 1.3 «Используемый инструментарий» посвящен программному обеспечению, применяемому в работе для реализации алгоритмов data mining, работы с машинным обучением и визуализации.

Программный комплекс был разработан на языке программирования Python. Прежде всего потому, что это хорошо документированный кросс-платформенный язык с понятным синтаксисом. В ходе работы использовались следующие библиотеки:

- Pandas — библиотека, обеспечивающая быструю и гибкую структуру данных. Данный программный модуль призван стать фундаментом для высокоуровневого практического анализа данных в Python.

Одномерная первичная структура данных Series и двумерная DataFrame, обрабатывают подавляющее большинство типичных случаев использования в финансах, статистике, социальных науках и во многих других областях.

Данный модуль используется в работе для идентификации и предварительной обработки биллингов.

- Для визуализации различных графиков и гистограмм частотных характеристик, полученных в результате обработки данных, используется инструмент Matplotlib.

Matplotlib — это библиотека Python для черчения в двумерном пространстве, которая используется в публикациях различных печатных форматов и интерактивных средах на разных платформах.

- В качестве инструмента для разработки интерфейса используется PyQt. PyQt — это набор «привязок» графического фреймворка Qt для языка программирования Python, выполненный в виде расширения Python.
- Для последующего после обработки данных этапа реализации алгоритма машинного обучения используются средства библиотеки Scikit-learn.
- Оценка построенной модели машинного обучения определяется с помощью инструментов библиотеки Statsmodels.

Statsmodels [6] — это модуль Python, который предоставляет классы и функции для оценки многих различных статистических моделей, а также для проведения статистических испытаний и исследования статистических данных.

Как и все описанные выше инструменты, программное обеспечение PyQt, применяемое в работе для отображения форм приложения, поддерживает концепцию open-source проекта, т. е. принцип создания ПО с открытым исходным кодом.

Язык Python и рассмотренные выше библиотеки подходят для решения задач из множества областей. В частности, эти инструменты предоставляют широчайшие возможности для реализации методов, позволяющих получить анализ биллингов, а именно частотные характеристики операций клиентов и межклиентские связи.

Это наглядным образом продемонстрировано во второй главе «Практическая часть», описывающей реализацию программного комплекса импорта и интеллектуального анализа данных об операциях.

Первый подраздел второй главы, именуемый «Используемые данные», посвящен структуре данных, на основе которой разрабатывались методы детализации биллингов. Для разработки методов детализации биллингов необхо-

дима соответствующая структура данных. Исходные данные можно разбить на две категории: непосредственно сам биллинг и данные о пользователе. К предполагаемым сведениям о человеке относятся его паспортные данные и номер телефона, который играет ключевую роль в анализе абонентских связей.

Так как в общем случае, детализация телефонных соединений является конфиденциальной информацией, для обработки и тестирования методов анализа данных, в этой работе используются данные сгенерированными случайным образом.

Так как вся информация о транзакциях хранится операторами связи в виде базы данных, то и сгенерированные данные, содержащие информацию о 100 пользователях и порядка нескольких тысяч операций, также записываются в базу данных. Затем сгенерированные данные могут быть экспортированы в приложение в поддерживаемом формате .csv. Теперь после запуска приложения, можно выбрать полученный файл и приступить к обработке данных.

После постановки проблемы и идентификации данных наступает этап предварительной обработки. Этому посвящен подраздел 2.2 «Используемые методы».

После загрузки файла и выбора телефонного номера, для которого необходим анализ, программа выводит данные в отсортированном по дате и времени виде.

Для обработанного файла с биллингами реализованы методы для получения таких данных, как информация о почасовой активности клиента за выбранные дни, частотные характеристики по произведенным операциям и визуализация звонков и сообщений в пределах одного дня.

Полный обзор разработанного программного продукта приведен во втором подразделе второй главы дипломной работы. В рамках данного раздела приводится поэтапное и тщательное описание работы с графическим интерфейсом программного продукта, снабженное большим количеством иллюстраций.

В подразделе 2.3 «Анализ связей между клиентами» решается задача классификации абонентов в зависимости от совершаемых ими операций, т. е. по наиболее частым связям с другими людьми. Для этого был реализован метод k-ближайших соседей из области машинного обучения с учителем. Так, на основе обучающей выборки абонентов с предопределенным кругом обще-

ния, функция классификатор определяет класс для всего множества клиентов.

Встроенная функция `accuracy_score` из библиотеки `Statsmodels` используется для определения точности. Точность как показатель помогает понять эффективность разработанного алгоритма.

Оценка точности оказалась равной 92.74110286. Таким образом на основе обучающей выборки было получено полное разбиение абонентов на круги общения.

В ходе анализа был выделен:

- класс `Close` людей, с которыми наиболее часто общался исследуемый клиент.
- класс `Middle` людей, имевший средний по частоте, но стабильный обмен звонками и смс сообщениями.
- класс `Outside` людей, с наименьшим числом вызовов и сообщений с испытуемым. Эта категория может включать случайные звонки и рекламный контент.

Для удобства полученные проанализированные данные представляются в виде рисунка, явно указывающего класс каждого клиента, и в виде таблицы с личными данными людей.

Исходя из полученных результатов можно делать выводы о круге абонентов, с которым общается выбранный для анализа человек.

Данная информация может быть полезна для предотвращения и раскрытия корпоративных преступлений, выявления преступных связей. Операторы сотовой связи могут использовать полученный анализ в маркетинговых целях, например, для выявления объема предоставляемых клиентам услуг.

В заключении содержатся основные результаты работы и краткие выводы по ним.

Работа завершается следующими приложениями с исходным кодом модулей программного продукта:

1. Приложение А. Листинг программы `main.py` — модуля с функциями импорта биллинга;
2. Приложение Б. Листинг программы `dialog.py` представляет собой модуль, со связующими функциями;
3. Приложение В. Листинг программы `incoming.py` — модуля, описывающего методы обработки и визуализации для входящих данных;

4. Приложение Г. Листинг программы `outcoming.py` представляет собой модуль, описывающий методы обработки и визуализации для исходящих данных;
5. Приложение Д. Листинг программы `classification.py` представляет собой реализацию модели машинного обучения;
6. Приложение Е. Листинг программы `data.py` представляет собой модуль для генерации тестовых данных;
7. Приложение Ж. Листинг программы `data.py` представляет собой модуль для генерации тестовых данных.

ЗАКЛЮЧЕНИЕ

Целью данной работы были разработка программного комплекса импорта данных, а также реализация методов, позволяющих получить анализ биллингов, в частности частотные характеристики операций клиентов и меж-клиентские связи.

В рамках реализации были разработаны методы первичного сбора статистики и её отображения в виде удобной для восприятия инфографики.

Для решения задачи классификации различных абонентов была реализована модель машинного обучения на основе алгоритма k-ближайших соседей.

На основе полученного анализа данных можно делать выводы о круге абонентов, с которым общается исследуемый человек, что в свою очередь оказывает бесценную помощь в задачах обеспечения безопасности.

Кроме того, обработанная информация и выявленные в ходе работы комплекса закономерности являются ценными данными для принятия бизнес-стратегий в области маркетинга.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Data mining [Электронный ресурс].— URL: <http://www.intuit.ru/studies/courses/6/6/lecture/158> (Дата обращения 12.05.2017). Загл. с экр. Яз. рус.
- 2 *Barber, D.* Bayesian Reasoning and Machine Learning / D. Barber. — Cambridge University Press, 2012.
- 3 *Царьков, С.* Алгоритм ближайшего соседа.— 2016.— URL: <https://basegroup.ru/community/articles/knn> (Дата обращения 13.05.2017). Загл. с экр. Яз. рус.
- 4 *Lantz, B.* Machine Learning with R: Second Edition / B. Lantz. — Packt Publishing Ltd., 2015.
- 5 Метод ближайших соседей.— 2015.— URL: <http://www.machinelearning.ru/wiki/index.php?title=KNN> (Дата обращения 04.05.2017). Загл. с экр. Яз. рус.
- 6 python statsmodels 0.8.0 documentation [Электронный ресурс].— URL: <https://pypi.python.org/pypi/statsmodels> (Дата обращения 5.06.2017). Загл. с экр. Яз. англ.

15.06.17 А. М. Мухомов