Министерство образования и науки Российской Федерации

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

Кафедра математической
кибернетики и компьютерных наук

СОЗДАНИЕ ПРИЛОЖЕНИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ С ПОМОЩЬЮ ОБЛАЧНОГО СЕРВИСА AZURE MACHINE LEARNING

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 09.03.04 — Программная инженерия
факультета КНиИТ
Илюшева Никиты Сергеевича

Научный руководитель
доцент, старш. преп. _____ М. И. Сафрончик

Заведующий кафедрой
к. ф.-м. н. _____ С. В. Миронов

Саратов 2018

ВВЕДЕНИЕ

В настоящее время сфера машинного обучения (Machine Learning) стремительно развивается и вызывает высокий интерес со стороны огромного количества компаний. Машинное обучение может быть описано как вычислительные системы, которые улучшаются с опытом. Со временем были успешно разработаны методы создания программных «моделей», которые обучаются на огромных объемах данных, а затем используются для прогнозирования определенных моделей, тенденций и результатов.

Относительно недавно Microsoft выпустила новый продукт Azure Machine Learning, позволяющий поднять решение задачи прогнозирования на совершенно новый уровень. Прогностическая аналитика - это базовая технология Azure Machine Learning, и ее можно просто определить как способ использования прошлого, чтобы предсказать будущее с целью помочь достичь желаемых результатов.

Цель бакалаврской работы — создание приложения интеллектуального анализа данных с помощью облачного сервиса Azure Machine Learning. Для ее достижения были поставлены следующие **задачи**:

1. выбрать подходящую для прогнозирования предметную область;
2. использовать Azure Machine Learning для создания и обучения модели;
3. использовать полученную модель для прогнозирования;
4. разместить веб-сервис и внедрить его в приложение;
5. протестировать работу приложения.

Методологические основы интеллектуального анализа данных представлены в работах Г. Чернышовой [2], А. Лунькова [3], Т. Афанасьевой [4] и Д. Барнса [6].

Структура и объем работы. Бакалаврская работа состоит из введения, 3-х разделов, заключения, списка использованных источников и 1-го приложения. Общий объем работы — 58 страниц, из них 40 страниц — основное содержание, включая 14 рисунков, список использованных источников информации — 18 наименований.

Первый раздел **«Интеллектуальный анализ данных»** посвящен описанию понятий интеллектуального анализа данных. В нем дается описание этапов построения модели интеллектуального анализа данных, рассматриваются задачи классификации, а также методы их решения, такие как де-

ревья решений и метод опорных векторов.

Второй раздел **«Облачный сервис Azure Machine Learning»** посвящен описанию облачного сервиса. В данном разделе описываются принципы контролируемого обучения, а также основные этапы высокоуровневого процесса Azure ML.

Третий раздел **«Создание решения прогнозной аналитики с использованием Azure Machine Learning»** посвящен процессам построения, обучения модели и разработки приложения. Представлен график, отображающий точность алгоритмов построения модели, а также скриншоты работы облачного сервиса и приложения.

1 Интеллектуальный анализ данных

Интеллектуальный анализ данных представляет собой процесс нахождения пригодных к использованию сведений в крупных наборах данных. В интеллектуальном анализе данных применяется математический анализ для выявления в больших объемах данных неочевидных, объективных и полезных на практике закономерностей. Процесс интеллектуального анализа данных напрямую связан с процессом принятия решений.

1.1 Этапы построения модели интеллектуального анализа данных

Процесс построения модели интеллектуального анализа данных можно представить как последовательность следующих базовых шагов:

— анализ предметной области;
— постановка задачи;
— подготовка данных;
— построение моделей;
— исследование и проверка моделей;
— развертывание и обновление моделей.

1.1.1 Анализ предметной области

В процессе изучения предметной области должна быть создана ее модель. Модель предметной области описывает процессы, происходящие в ней, и данные, которые используются в этих процессах. От того, насколько верно смоделирована предметная область, зависит успех дальнейшей разработки приложения интеллектуального анализа данных.

1.1.2 Постановка задачи

Постановка задачи интеллектуального анализа включает в себя формулировку и формализацию задачи [1].

1.1.3 Подготовка данных

Происходит определение и анализ требований к данным. Затем происходит сбор данных. Источником являются хранилища данных, оперативные, справочные и архивные базы данных. Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки [2].

Для обеспечения качественного анализа необходимо проведение предварительной обработки данных. Данные, полученные в результате сбора, должны соответствовать определенным критериям качества. При наличии данных низкого качества необходимо проводить очистку данных [3].

1.1.4 Построение моделей

Происходит создание структуры интеллектуального анализа данных и ее последующая обработка. Обработка структуры называется обучением и подразумевает применение алгоритма интеллектуального анализа данных с целью выявить искомые закономерности. В ходе построения модели используется набор данных для обучения.

1.1.5 Исследование и проверка моделей

Проверяется эффективность работы модели. Кроме того, во время построения модели обычно создается несколько моделей с различной конфигурацией, а затем проверяются все модели, чтобы определить, какая из них обеспечивает лучшие результаты для поставленной задачи и имеющихся данных. Для проверки точности модели набор проверочных данных.

1.1.6 Развертывание и обновление моделей

Эффективные модели могут быть применимы для решения разного рода задач и принятия решений в рассматриваемой сфере деятельности человека.

1.2 Задачи интеллектуального анализа данных

В интеллектуальном анализе данных можно выделить несколько вариантов классификаций задач.

По типам производимой информации классифицируются на задачи: классификации, кластеризации, прогнозирования, ассоциации и визуализации.

По стратегиям задачи разделяются на следующие группы: обучение с учителем и обучение без учителя.

В зависимости от используемых моделей задачи делятся на описательные и прогнозирующие.

1.3 Задачи классификации и прогнозирования

Классификация – это процесс распределения по определенному принципу множества предметов, процессов или явлений, имеющих признаки для определения сходства или различия между ними.

Процесс классификации состоит из двух этапов: построения модели и ее использования. Перед построением происходит разделение на обучающее и тестовое множества путем деления выборки в определенной пропорции. Использование модели заключается в классификации новых или неизвестных значений. Известные значения из тестового примера сравниваются с результатами использования полученной модели. Процент правильно классифицированных примеров в тестовом множестве демонстрирует точность модели [4].

Прогнозирование – установление функциональной зависимости между зависимыми и независимыми переменными. Целью прогнозирования является предсказание будущих событий.

Различие задач классификации и прогнозирования состоит в том, что в первой задаче предсказывается класс зависимой переменной, а во второй – числовые значения зависимой переменной.

1.4 Деревья решений

Деревья решений - это метод, позволяющий предсказывать принадлежность объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких предикторных переменных [2].

В ходе процесса построения дерева алгоритм должен найти критерий разбиения. Каждый узел проверки должен быть отмечен определенным атрибутом, который должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса или же были максимально приближены к такому разбиению.

В алгоритме CART реализован критерий расщепления, который называется индексом Gini. При помощи этого индекса атрибут выбирается на основании расстояний между распределениями классов. Если дано множество T, включающее примеры из n классов, индекс Gini определяется по формуле

$$Gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

, где T – текущий узел; p_j – вероятность класса j в узле T; n – количество классов.

В процессе построения дерева, чтобы его размеры не стали чрезмерно большими, используют специальные процедуры, такие как сокращение дерева путем отсечения ветвей и использование правил остановки обучения [5].

1.5 Метод опорных векторов

Метод опорных векторов (Support Vector Machine – SVM) относится к группе граничных методов, которая определяет классы при помощи границ областей. Цель метода опорных векторов – найти плоскость, разделяющую два множества объектов. Метод отыскивает объекты, лежащие на границах областей. Классификация считается хорошей, если область между границами пуста.

Один из вариантов метода опорных векторов – линейный SVM. Данный метод заключается в поиске некоторой линейной функции f(x), принимающей значения меньше нуля для векторов одного класса и больше нуля – для векторов другого. В качестве исходных данных для поиска классифицирующей функции f(x) дан тренировочный набор векторов пространства, для которых известна их принадлежность к одному из классов. Гиперплоскость может быть задана формулой

$$f(x) = ax + b$$

, где f – вектор к разделяющей гиперплоскости; b – вспомогательный параметр; x – объект.

В результате решения задачи найдена функция, при помощи которой для каждого нового объекта отрицательное или положительное значение определяет принадлежность объекта к одному из классов [2].

2 Облачный сервис Azure Machine Learning

2.1 Общие сведения

Azure Machine Learning - полностью управляемая платформа для машинного обучения, которая позволяет выполнять задачи прогнозной аналитики. Данный сервис позволяет создавать модели и интегрировать их в промышленные решения. Разработка моделей (экспериментов) достигается с помощью Azure ML Studio, веб-среды проектирования, доступной через браузер и предоставляющей визуальный интерфейс для создания моделей в стиле `Drag\&Drop`.

Одной из наиболее важных функций Azure ML является ее служба публикации прогнозной модели в виде веб-сервиса. Он получает на вход данные, по которым пользователь желает получить прогноз, обрабатывает их, и на выходе пользователь получает интересующие его результаты прогноза. Законченный (обученный) эксперимент может быть представлен как веб-API, который может быть использован любым другим приложением, таким как веб-сайт, мобильное приложение и так далее.

2.2 Высокоуровневый рабочий процесс Azure Machine Learning

Основной процесс создания решений Azure Machine Learning состоит из повторяемого шаблона шагов рабочего процесса, которые разработаны для быстрого создания нового интеллектуального аналитического решения. Основные этапы процесса:

- Данные.
- Создание модели.
- Оценка модели.
- Уточнение и оценка модели.
- Развертывание модели.
- Тестирование и использование модели.

Одно из первых основных различий в понимании Azure Machine Learning - это концепция контролируемого и неконтролируемого обучения [6].

2.3 Контролируемое обучение

Контролируемое обучение - это тип алгоритма машинного обучения, который использует известные наборы данных для создания модели, которая

затем может делать прогнозы. Известные наборы данных вызывают и включают элементы входных данных вместе с известными значениями ответа. Из этих тренировочных наборов контролируемые алгоритмы обучения пытаются построить новую модель, которая может делать прогнозы на основе новых входных значений вместе с известными результатами.

Использование контролируемого подхода к обучению для создания новых прогнозирующих моделей требует учебных наборов данных. После создания новая модель может быть затем проверена на точность с использованием тестовых наборов данных.

Процесс оценки новых моделей прогнозирования, который использует контролируемое обучение, в первую очередь состоит в определении точности новой сгенерированной модели. В этом случае точность предсказательной модели может быть легко определена, поскольку входные значения и результаты уже известны.

2.4 Развертывание модели прогнозирования

Развертывание новой модели прогнозирования принимает форму публикации веб-службы в Интернете через Microsoft Azure. Затем веб-службу можно вызвать через протокол передачи состояния представления (REST).

Когда в Интернете отображается новая модель прогнозирования машинного обучения, она выполняет следующие операции:

1. Новые входные данные передаются в веб-службу в виде полезной нагрузки объекта JavaScript Object Notation (JSON).
2. Затем веб-служба передает входящие данные в качестве входных данных в модель прогнозирования Azure Machine Learning.
3. Затем модель обучения Azure Machine Learning генерирует новое предсказание на основе входных данных и возвращает новые результаты предсказания вызывающему абоненту через полезную нагрузку JSON [6].

3 Создание решения прогнозной аналитики с использованием Azure Machine Learning

В данной работе проводится оценка вероятности того, выиграет ли команда футбольный матч при определенных условиях. В качестве входных данных будет использоваться набор данных о прошедших футбольных матчах. С использованием информации в наборе будет создана прогностическая модель. Затем, после развертывании веб-сервиса можно будет предсказывать вероятность победы команды в матче на основе новых введенных данных.

3.1 Подготовка данных

В данном эксперименте данные импортируются путем загрузки файла формата CSV. В этом наборе содержится информация о девяти атрибутах. Для включения всех атрибутов в данный набор была изучена информация о том, какие факторы влияют на победу или поражение команды в футбольном матче.

3.2 Создание и обучение модели

Для прогнозирования значений выбран атрибут `Chance to Win`, который может принимать значение `Low` (низкий шанс на победу) или `High` (высокий шанс).

Далее разрабатывается модель на основе данных, с использованием одного из встроенных алгоритмов в Azure ML. Для этого данные разделяются на два случайных набора, при помощи инструмента `Split Data`. 80% данных используется для обучения и 20% отведено для оценки:

— Данные обучения. Эта группа используется для создания новой предсказательной модели на основе исторических данных, с помощью алгоритма ML, который используется для решения.

— Данные оценки. Эта группировка используется для тестирования новой модели прогнозирования по сравнению с известными результатами для определения точности и вероятностей [7].

В данном эксперименте для обучения модели применяются алгоритмы (`Two-Class Boosted Decision Tree`, `Two-Class Support Vector Machine`), а также инструмент `Train Model`. Это делается с целью продемонстрировать работу алгоритмов, сравнить полученные результаты и выбрать из них для

дальнейшего развертывания и прогнозирования один, дающий наилучшие результаты точности.

3.3 Оценка модели

На этом этапе при помощи инструмента `Score Model` можно увидеть результаты оценки модели. В наборе появляются два дополнительных атрибута, которые отображаются в полученном наборе данных: оценочные метки и оценочные вероятности. Эти новые столбцы в наборе данных представляют собой то, что, помимо расчета прогноза для каждой строки, алгоритм также может обеспечивать численный коэффициент вероятности. Этот фактор вероятности представляет собой потенциал модели для точного прогнозирования каждой строки в наборе данных на основе конкретных значений, найденных в каждом из других столбцов строки.

3.4 Уточнение и оценка модели

Создается набор кривых и показателей, которые позволяют просматривать или сравнивать результаты двух моделей. График `ROC` отображает долю истинных положительных результатов из всех фактических положительных результатов. Диагональная линия соответствует 50-процентной точности в прогнозах и может использоваться в качестве эталона, который можно улучшить. Чем выше и дальше влево уходит кривая, тем точнее модель. Согласно графику можно сделать вывод, что алгоритм, основанный на построении дерева решений, лучше, так как дает предсказания с большей точностью.

3.5 Развертывание модели

После публикации веб-сервиса появляется панель инструментов, содержащая ключи API и справочные ссылки API для нового веб-сервиса прогнозируемой модели. Панель управления Azure ML Studio предоставляет всю информацию, необходимую для вызова новой модели прогнозирования через Интернет. Также можно увидеть ключ API - это уникальный идентификатор безопасности, который необходимо передать каждому запросу веб-службы для аутентификации вызывающего приложения.

3.6 Тестирование веб-сервиса

На этом этапе выполняются самые простые действия. Необходимо ввести значения, которые будут использоваться для вызова веб-службы и для

которых необходимо получить прогноз, а затем просмотреть результирующий ответ. При тестировании можно увидеть поля для ввода данных и выходную информацию вместе с полученными результатами.

3.7 Создание приложения

При разработке приложения и внедрении в него веб-сервиса использовалось программное обеспечение `Microsoft Visual Studio`. Разрабатываемым приложением является веб-страница, которая предоставляет возможность ввода данных, а также после обращения к веб-сервису предоставляет желаемый результат. При разработке используется язык C#.

В состав Visual Studio входит инструмент IIS Express, представляющий собой усеченную версию сервера приложений Microsoft, предназначенную для запуска разрабатываемых приложений ASP.NET. Когда приложение запускается из Visual Studio, сервер IIS Express начинает свою работу и прослушивает входящие запросы (на порте). Как только сервер IIS Express запускается, Visual Studio создает новое окно браузера и применяет его для перехода на URL, который приводит к загрузке файла `WebForm.aspx` из IIS Express [8].

Создаваемая веб-страница разделяется на две части: веб-форму и файл, содержащий программный код. При этом форма сохраняется в файле с расширением ASPX, а программный код – с расширением CS. Такая модель обеспечивает лучшую организацию элементов веб-приложения за счет отделения пользовательского интерфейса от программной логики [9].

При работе приложения происходит обращение к веб-сервису, доступ к которому осуществляется при помощи API-ключа, указанного в программном коде. Результат возвращается в виде значений всех атрибутов в наборе данных, а также двух дополнительных значений, который демонстрируют шанс на победу команды в матче.

ЗАКЛЮЧЕНИЕ

В ходе работы изучены основные понятия интеллектуального анализа данных, а также принципы работы с облачным сервисом Azure Machine Learning. На основе выбранной предметной области с использованием алгоритмов машинного обучения, внедренных в сервис, была построена и обучена модель для прогнозирования шансов на победу в футбольном матче. Полученные результаты опубликованы в виде веб-сервиса, который был использован для разработки веб-приложения на языке C# в среде Microsoft Visual Studio.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 НОУ ИНТУИТ | Лекция | Процесс Data Mining. Начальные этапы [Электронный ресурс]. — URL: https://www.intuit.ru/studies/courses/6/6/lecture/192 (Дата обращения 03.06.2018). Загл. с экр. Яз. рус.

2 *Чернышова, Г.* Интеллектуальный анализ данных: учеб.пособие для студентов специальности 080801.65 «Прикладная информатика (в экономике)» / Г. Чернышова. — Саратов: Саратовский государственный социально-экономический университет, 2012.

3 *Луньков, А.* Интеллектуальный анализ данных. Часть 1: учеб.-метод. пособие / А. Луньков, А. Харламов. — Саратов: СГУ им. Н.Г. Чернышевского.

4 *Афанасьева, Т.* Введение в проектирование систем интеллектуального анализа данных: учебное пособие / Т. Афанасьева, А. Афанасьев. — Ульяновск: УлГТУ, 2017.

5 Преимущества использования деревьев решений - Метод построения дерева решений [Электронный ресурс]. — URL: https://studwood.ru/1012389/menedzhment/preimuschestva_ispolzovaniya_derevev_resheniy (Дата обращения 06.06.2018). Загл. с экр. Яз. рус.

6 *Barnes, J.* Azure Machine Learning Microsoft Azure Essentials / J. Barnes. — Redmond: Microsoft Press, 2015.

7 Azure Machine Learning Tutorial using Python, API and Excel – Microsoft Faculty Connection [Электронный ресурс]. — URL: https://blogs.msdn.microsoft.com/uk_faculty_connection/2016/10/06/azure-machine-learning-tutorial-using-python-api-and-excel/ (Дата обращения 02.06.2018). Загл. с экр. Яз. англ.

8 ASP.NET Web Forms 4.5 [Электронный ресурс]. — URL: https://professorweb.ru/my/ASP_NET/webforms_4_5/level1/1_0.php (Дата обращения 06.06.2018). Загл. с экр. Яз. рус.

9 Создание Web-приложений средствами ASP.NET [Электронный ресурс]. — URL: http://5fan.ru/wievjob.php?id=51475 (Дата обращения 06.06.2018). Загл. с экр. Яз. рус.