ОПТИМИЗАЦИЯ ПРОЦЕССОВ ПЛАНИРОВАНИЯ ЗАПРОСОВ БАЗ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Калинина Е.С.[1], Манохина Т. В.[2], Ступаков С. А.[3]

[123]ФГБОУ ВО «Омский государственный университет путей сообщения», Омск, Россия

В данной статье рассмотрена проблематика оптимизации процессов планирования запросов баз данных методами машинного обучения. Целью статьи является исследование средств решения задач в Machine Learning с помощью SQL Server Machine Learning Services для определения наиболее эффективного метода их программной реализации с использованием встроенных языковых средств SQL Server или классического способа обработки данных. Была осуществлена оптимизация плана выполнения запроса с использованием средств машинного обучения позволяющий увеличения производительности выполнения запросов. С целью уменьшений время его обработки. Проведена разработка специального математического и программного обеспечения системы управления схемой реляционной базы данных с использованием методов машинного обучения для ускорения обработки запросов. Новизна исследования – определение количественной оценки эффективности применения встроенных языковых средств SQL Server Machine Learning Services. Практическая ценность работы – результаты исследований могут быть внедрены во многих сферах: производство, транспортные системы, медицина, образование и т.д.

Введение

С каждым годом развитие информационных технологий во всех сферах деятельности человечества вызывает необходимость поиска и разработки новых способов использования Интернета и его возможностей. Это устойчивое и последовательное движение к построению информационного общества и является следствием интенсивного развития компьютерных и

телекоммуникационных технологий [1]. Технологии машинного обучения все активнее проникают в повседневную жизнь, и мы даже не задумываемся о том, что нашу ленту в Instagram и других социальных сетях сформировал именно искусственный интеллект. Конечно, у него есть и более серьезные задачи – например, прогноз спроса на товары, распознавание лиц, отпечатков или голоса [2].

Исследование методов программной реализации машинного обучения и эффективной обработки данных является достаточно актуальной темой, поскольку потребность в использовании компьютерной техники, программных продуктов высокого качества с каждым днем растет все больше. Применение машинного обучения может значительно ускорить и повысить эффективность принятия решений, прогнозировать и анализировать поведение системы, благодаря чему можно будет работать в режиме реального времени и избежать нежелательных ситуаций [3, 4].

Методология исследования

Целью статьи является исследование средств решения задач в Machine Learning с помощью SQL Server Machine Learning Services для определения наиболее эффективного метода их программной реализации: с использованием встроенных языковых средств SQL Server или классического способа обработки данных.

Объект исследования – процессы решения задач в Machine Learning. Предмет исследования – методы и алгоритмы решения задач машинного обучения с использованием возможностей SQL Server Machine Learning Services. Методы исследования – проведение экспериментов по эффективности применения средств решения задач в Machine Learning с помощью SQL Server Machine Learning Services и статистическая обработка полученных результатов. Для проведения экспериментов по эффективности обработки запросов, данных использовалась база данных, содержащая информацию о количестве прокатов лыж [5].

Результаты и обсуждение

Прокат лыж имеет сезонный характер, поэтому необходимо правильно сгруппировать данные таблицы rental_data, чтобы в дальнейшем их анализировать. Сезон начинается с декабря месяца предыдущего года и заканчивается в апреле текущего года. Как уже отмечалось выше, для проведения исследований необходимо создать таблицу с большим объемом данных, чем мы имеем. Также было бы хорошо иметь возможность регулировать количество сгенерированных данных по некоторому показателю. Формула для генерации количества аренд лыжного снаряжения выглядит следующим образом [6]:

$RentalCount$ = 1000 ∗ $MonthCoeficient$ ∗ EXP (−0.07 ∗ Текущий день) (1)

1000 – базовое количество аренд, которое компания может обеспечить. В формуле используется функция EXP. Отрицательный показатель степени (-0,07) характеризует спад функции.

Для обеспечения уникальности данных был рассчитан дополнительный коэффициент (MonthCoeficient) по следующей формуле [7]:

- если текущий месяц – декабрь:

$$MonthCoeficient = 1 - \text{Номер\_месяца} / 100 * 0.5 \quad (2)$$

- если текущий месяц – январь, февраль, март или апрель:

$$MonthCoeficient = 1 - \text{Номер\_месяца} / 100 * \text{Номер\_месяца} \quad (3)$$

Такое поколение коэффициентов регулирует то, как номер месяца влияет на количество аренд. Да, в январе коэффициент будет наибольшим, поскольку именно в этом месяце наблюдается наибольший спрос на лыжное снаряжение. Кроме месяца на результат также могут влиять дополнительные факторы, способствующие росту количества аренд: шел ли снег в этот день (если да – количество увеличивается на 20); тип дня (праздничный, выходной, будний). Для праздничного дня – 20 дополнительных аренд, для выходного – 30. Данные о том, шел ли снег генерировались случайным образом по следующему условию: если текущий месяц – январь, февраль или декабрь – @Show = ROUND (RAND () * (1 - 0), 0). В противном случае @Show=0.

Для создания большого количества данных была использована сохраненная процедура с параметром. Параметр определяет количество итераций, в течение которых будет происходить генерация исходных данных. Скрипт включает все условия, изложенные в математический постановке. Диапазон входных данных – с 01/01/2000 по 12/11/2022. Приведенный алгоритм является удобным, поскольку можно с легкостью корректировать количество сгенерированных данных путем редактирования переменной @NumberOfIterations. Для каждой даты в цикле осуществляется подсчет количества аренд лыжного снаряжения. Поскольку генерация результата могла проводиться несколько раз для той же даты, то использовалась функция RAND () для предотвращения дубликатов. В результате была сформирована таблица rental_data_new_exp, содержащая 484 000 строк. Генерация данных производилась довольно быстро (до 10 минут для наибольшего объема данных) [8].

Далее опишем описание результатов моделирования. Наиболее популярным алгоритмом машинного обучения является линейная регрессия благодаря простоте и скорости осуществления прогнозирования. Поэтому обработка данных при проведении опытов будет происходить с использованием этого метода. Проведем исследования, чтобы определить, как количество данных влияет на время их обработки различными способами: с использованием встроенных языковых средств SQL Server и классического подхода. Поскольку эксперименты будут проводиться с использованием серверов, имеющих разное расположение, то в табл. 1 представлены строки подключения к ним. Свойства локально расположенного сервера и сервера в Microsoft Azure представлены и в табл. 2 [10].

Табл. 1. Подключение к серверу БД

Сервера	Строка подключения
Локальный	DRIVER={ODBC Driver 13 для SQL Server}; SERVER=DESKTOPA6FSIKB\SQLEXPRESS; DATABASE=TutorialDB; Trusted_Connection=yes;

В облачной платформе Microsoft Azure	Driver={ODBC Driver 13 for SQL Server};Server= tcp:23.96.37.83,1401;Database=TutorialDB;Uid=anna_koskina;Pwd=******; En crypt=yes;TrustServerCertificate=no;Connection Timeout= 30;

Табл. 2. Свойства серверов

Свойство	Локальный сервер	Сервер виртуальной машины в Microsoft Azure
Версия	14.0.2027.2	14.0.3356.20
Количество процессоров	2	2
Объем физической памяти	4020 MB	8192 MB
Объем виртуальной памяти	131 071 GB	128 000 GB

Для анализа времени обработки данных на локальном сервере используем функцию библиотеки Python timeit default_timer(), при работе с сервером Microsoft Azure будем использовать функцию R Sys.time(). Основными процессами, которые взаимодействуют с данными, являются их загрузка, обработка (обучение модели и прогнозирование результатов) и пересылка. Для доступа к данным необходимо указать строку подключения, в которой указать имя сервера, базу данных и таблицу. Проведение процесса машинного обучения требует определения выборки, на основе которой будет проводиться обучение модели и другой – для прогнозирования результатов. Итак, поскольку в качестве входных данных использовалась информация за 2000-2015 годы, то обучение модели проводилось на основе данных за 2000-2014 годы, прогнозирование – для 2015 года. Программная реализация работы с данным при помощи классического способа представлена на рис. 1. Обработка данных на сервере производилась с использованием встроенных языковых средств SQL Server. Для этого выполнялась сохраненная процедура, текст которой представлен на рис. 2 и 3.

```
import timeit
import pyodbc
import pandas
from sklearn.linear_model import LinearRegression
# Connection string to your SQL Server instance
conn_str = pyodbc.connect("Driver={ODBC Driver 13 for SQL
Server};Server=23.96.37.83,1401;Database=TutorialDB;Uid=anna_koskina;Pwd=********;
Encrypt=yes;TrustServerCertificate=no;Connection Timeout=30;")
query_str = "SELECT  Year, Month, Day, Rentalcount, Weekday, Holiday, Snow FROM
dbo.rental data new exp"
```

Рис. 4. Программная реализация применения метода машинного обучение для обработки данных

```
CREATE PROCEDURE [dbo].[LinearRegressionPrediction]
AS
BEGIN
    EXECUTE sp_execute_external_script
      @language = N'Python'
    , @script = N'
from sklearn.linear_model import LinearRegression
import timeit
```

Рис. 5. Текст сохраненной процедуры, которая выполняет обработку данных на сервере, используя метод машинного обучения (язык программирования Python)

```
CREATE PROCEDURE [dbo].[LinearRegressionPredictionR]
AS
BEGIN
EXECUTE sp_execute_external_script
      @language = N'R'
    , @script = N'
```

Рис. 6. Текст сохраненной процедуры, которая выполняет обработку данных на сервере используя метод машинного обучения (язык программирования R)

Скрипты для проведения исследований эффективности работы встроенных языковых средств SQL Server представлены на рис. 7-8.

```
import pyodbc
import timeit
# Connection string to your SQL Server instance
conn_str = pyodbc.connect("DRIVER={ODBC Driver 13 for SQL Server}; SERVER=DESKTOP-
A6FSIKB\SQLEXPRESS; DATABASE=TutorialDB; Trusted_Connection=yes")

cursor = conn_str.cursor()
cursor.execute("CHECKPOINT;")
cursor.execute("DBCC FREEPROCCACHE;")
cursor.execute("DBCC DROPCLEANBUFFERS;")
cursor.execute("EXEC LinearRegressionPrediction;")
# Calculate the time required for transfering data from SQL Server
# startDataTransfering - capture the beginning of data transfering
# endDataTransfering - capture the end of data transfering
startDataTransfering = timeit.default_timer()
rc = cursor.fetchall()
endDataTransfering = timeit.default_timer()
print(rc)

print('Time data transfering: ', endDataTransfering - startDataTransfering)
cursor.close()
conn_str.close()
```

Рис. 7. Исследование эффективности работы встроенных языковых средств SQL Server (локальный сервер)

```
# Calculate the time required for transfering data from SQL Server
# startDataTransfering - capture the beginning of data transfering
# endDataTransfering - capture the end of data transfering

startDataTransfering  <- Sys.time()

res <- dbFetch(query)
tail(res, 2)

endDataTransfering <- Sys.time()

dataTransfering <- endDataTransfering - startDataTransfering

dataTransfering
```

Рис. 8. Исследование эффективности работы встроенных языковых средств SQL Server (сервер Microsoft Azure)

Анализируя все полученные результаты можно заключить, что обработка данных на сервере средствами SQL Server Machine Learning Services оказалась значительно более эффективной, чем использование классического подхода к работе с данными.

Такой результат обуславливается следующими причинами:

- не тратилось дополнительно время на загрузку данных (поскольку работа с данными происходила на том же сервере, где они хранятся); процесс обработки данных путем использования метода машинного обучения происходил значительно быстрее по сравнению с классическим способом;

- SQL Server предоставляет возможность создания сохраненных процедур с исходным кодом Python, что способствует повышению производительности работы системы, позволяет многократное использование процедур разными пользователями, что является более безопасным; возможен запуск параллельных работ на сервере для повышения эффективности работы с данными (параметр parallel=1 команды sp_execute_external_script).

Также было обнаружено, что эффективность применения SQL Server Machine Learning Services становилась более значимой при увеличении данных, которые использовались для исследования, что является весомым аргументом, поскольку обычно для анализа используются миллионы записей, которые нужно быстро обработать и показать результат. Ускорить процесс обработки данных на сервере можно, отредактировав определенные характеристики сервера – увеличить количество процессоров, объем физической и виртуальной памяти, использовать последние версии SQL Server.

Заключение

Путем проведения экспериментов было определено, что применение SQL Server ML Services позволяет ускорить процесс обработки данных в 2 – 4 раза для локально расположенного сервера и в 2.5 раза для находящегося в облачной платформе сервера. Выявлено, что при увеличении количества данных эффективность использования встроенных языковых средств SQL Server становилась более заметной. Определены эффективность применения сервера БД для обработки данных и преимущества данного способа. Было обнаружено, что задачи исследования решаются в 2 – 4 раза быстрее, чем при классическом способе обработки данных, что позволяет работать в режиме реального времени. Благодаря этому можно мгновенно реагировать на изменения и избегать нежелательных последствий работы системы. Определен наиболее эффективный способ обработки данных, использование которого позволяет уменьшить время, необходимое для получения конечного результата решения задачи средствами машинного обучения.

Литература

1. Аксютина Е. М., Белов Ю. С. Обзор архитектур и методов машинного обучения для анализа больших данных // Электронный журнал: наука, техника и образование. - 2016. №1 (5). - С. 132–139.

2. Белов Ю. С., Козина А. В., Гришунов С. С. Применение критерия «сигнал/шум» для определения эффективности методов машинного обучения // Известия ТулГУ. Технические науки. - 2018. - № 12. - С. 292–295.

3. Боровский А. А. Перспективы применения технологий машинного обучения к обработке больших массивов исторических данных // Кибернетика и программирование. – 2015. – № 1. – С. 77 – 114.

4. Гусев А. В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения // Искусственный интеллект в здравоохранении. – 2017. - №3. - С. 92–105.

5. Жуков Д. А., Клячкин В. Н. Задачи обеспечения эффективности машинного обучения при диагностике технических объектов // Электронный журнал: Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. - 2016. - № 10. - С. 172–174.

6. Как Big Data с Machine Learning борются с пробками и улучшают дороги. – URL: https://www.bigdataschool.ru/blog/big-data-machine-learning-iot-transporttraffic.html.81 (дата обращения: 17.11.2022).

7. Кафтанников И. Л., Парасич А. В. – Проблемы формирования обучающей выборки в задачах машинного обучения // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». - 2016. - № 3. - С. 15–24.

8. Кондрашов Ю. Н. Анализ данных и машинное обучение на платформе MS SQL Server. – URL: https://aldebaran.ru/author/n_kondrashov_yu/kniga_analiz_dannyih_i_mashinnoe_obuchenie_na_ (дата обращения: 17.11.2022).

9. Коротеев М. В. Обзор некоторых современных тенденций в технологии машинного обучения // Технологии искусственного интеллекта в менеджменте. – 2018. - № 1. - С. 26–35.

10. Краснянский М. Н., Обухов А. Д., Соломатина Е. М., Воякина А. А., – Сравнительный анализ методов машинного обучения для решения задачи классификации документов научно-образовательного учреждения – 2018 // Вестник ВГУ, серия: системный анализ и информационные технологии. – 2018. - № 3. - С. 1038–1082.

References in Cyrillics

1. Aksjutina E. M., Belov Ju. S. Obzor arhitektur i metodov mashinnogo obuchenija dlja analiza bol'shih dannyh // Jelektronnyj zhurnal: nauka, tehnika i obrazovanie. - 2016. №1 (5). - S. 132–139.

2. Belov Ju. S., Kozina A. V., Grishunov S. S. Primenenie kriterija «signal/shum» dlja opredelenija jeffektivnosti metodov mashinnogo obuchenija // Izvestija TulGU. Tehnicheskie nauki. - 2018. - № 12. - S. 292–295.

3. Borovskij A. A. Perspektivy primenenija tehnologij mashinnogo obuchenija k obrabotke bol'shih massivov istoricheskih dannyh // Kibernetika i programmirovanie. – 2015. – № 1. – S. 77 – 114.

4. Gusev A. V. Perspektivy nejronnyh setej i glubokogo mashinnogo obuchenija v sozdanii reshenij dlja zdravoohranenija // Iskusstvennyj intellekt v zdravoohranenii. – 2017. - №3. - S. 92–105.

5. Zhukov D. A., Kljachkin V. N. Zadachi obespechenija jeffektivnosti mashinnogo obuchenija pri diagnostike tehnicheskih ob#ektov // Jelektronnyj zhurnal: Sovremennye problemy proektirovanija, proizvodstva i jekspluatacii radiotehnicheskih sistem. - 2016. - № 10. - S. 172–174.

6. Kak Big Data s Machine Learning borjutsja s probkami i uluchshajut dorogi. – URL: https://www.bigdataschool.ru/blog/big-data-machine-learning-iot-transporttraffic.html.81 (data obrashhenija: 17.11.2022).

7. Kaftannikov I. L., Parasich A. V. – Problemy formirovanija obuchajushhej vyborki v zadachah mashinnogo obuchenija // Vestnik JuUrGU. Serija «Komp'juternye tehnologii, upravlenie, radiojelektronika». - 2016. - T. 16, № 3. - S. 15–24.

8. Kondrashov Ju. N. Analiz dannyh i mashinnoe obuchenie na platforme MS SQL Server. – URL: https://aldebaran.ru/author/n_kondrashov_yu/kniga_analiz_dannyih_i_mashinnoe_obuchenie_na_ (data obrashhenija: 17.11.2022).

9. Koroteev M. V. Obzor nekotoryh sovremennyh tendencij v tehnologii mashinnogo obuchenija // Tehnologii iskusstvennogo intellekta v menedzhmente. – 2018. - № 1. - S. 26–35.

10. Krasnjanskij M. N., Obuhov A. D., Solomatina E. M., Vojakina A. A., – Sravnitel'nyj analiz metodov mashinnogo obuchenija dlja reshenija zadachi klassifikacii dokumentov nauchno-obrazovatel'nogo uchrezhdenija – 2018 // Vestnik VGU, serija: sistemnyj analiz i informacionnye tehnologii. – 2018. - № 3. - S. 1038–1082.

Ключевые слова

Оптимизация, планирование, запросы, базы данных, машинное обучение

Калинина Екатерина Сергеевна, к. т. н, доцент кафедры «Информатика и компьютерная графика» ФГБОУ ВО «Омский государственный университет путей сообщения», Омск, Россия,
ekkalinina@mail.ru

Манохина Татьяна Витальевна, старший преподаватель кафедры «Информатика и компьютерная графика» ФГБОУ ВО «Омский государственный университет путей сообщения»,
mtv-gups@mail.ru

Ступаков Сергей Анатольевич, к. т. н, доцент кафедры «Информатика и компьютерная графика» ФГБОУ ВО «Омский государственный университет путей сообщения», Омск, Россия,
stupakov1@yandex.ru

Keywords
Optimization, planning, queries, databases, machine learning

Kalinina Ekaterina Sergeevna, Candidate of Technical Sciences, Associate Professor of the Department of Informatics and Computer Graphics, Omsk State Transport University, Omsk, Russia,
ekkalinina@mail.ru

Manokhina Tatyana Vitalievna, Senior Lecturer of the Department of Informatics and Computer Graphics, Omsk State Transport University,
mtv-gups@mail.ru

Stupakov Sergey Anatolyevich, Candidate of Technical Sciences, Associate Professor of the Department of Informatics and Computer Graphics, Omsk State Transport University, Omsk, Russia,
stupakov1@yandex.ru

OPTIMIZATION OF BASE QUERY PLANNING PROCESSES BY MACHINE LEARNING METHODS

The article deals with the problem of optimizing database query planning processes using machine learning methods. The purpose of the article is to explore the tools for solving problems in Machine Learning using SQL Server Machine Learning Services to determine the most effective method for their programmatic implementation using the built-in language tools of SQL Server or the classical way of data processing. The query execution plan was optimized using machine learning

tools to increase the performance of query execution in order to reduce its processing time. The development of a special mathematical and software system for managing a relational database schema using machine learning methods to speed up query processing. The novelty of the research is the determination of a quantitative assessment of the effectiveness of using the built-in language tools of SQL Server Machine Learning Services. The practical value of the work - the results of research can be implemented in many areas: production, transport systems, medicine, education, etc.