Компьютерные инструменты в образовании, 2024
№ 2: 58–71
УДК: 004.65
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-2-58-71

РЕАЛИЗАЦИЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ СРЕДСТВАМИ РЕЛЯЦИОННЫХ СУБД

Графеева Н. Г.[1], канд. физ.-мат. наук, nggrafeeva@corp.ifmo.ru
Назаров А. А.[1], студент, ✉ artem.a.nazarov@yandex.ru

[1]Национальный исследовательский университет ИТМО,
Кронверкский пр., 49, лит. А, 197101, Санкт-Петербург, Россия

Аннотация

В статье рассматривается проблематика интеграции алгоритмов машинного обучения в реляционные СУБД. Проведен обзор и сравнительный анализ текущих технических возможностей реляционных СУБД Oracle, PostgreSQL, SQL Server, DB2 и MySQL, адаптированных для интеллектуального анализа данных. На основе полученных результатов сделаны выводы об уровне готовности современных СУБД к решению задач анализа данных.

Ключевые слова: *реляционные средства управления базами данных, интеллектуальный анализ данных, технические решения.*

Цитирование: Графеева Н. Г., Назаров А. А. Реализация алгоритмов машинного обучения средствами реляционных СУБД // Компьютерные инструменты в образовании. 2024. № 2. С. 58–71. doi:10.32603/2071-2340-2024-2-58-71

1. ВВЕДЕНИЕ

Постоянно растущие объемы информации ставят перед нами задачи ее хранения и обработки. Для решения первой задачи в большинстве технических проектов продолжают использоваться реляционные системы управления базами данных (далее — СУБД) сразу по нескольким причинам: соответствие требованиям ACID (atomicity, consistency, isolation, durability), наличие развитой теории реляционной модели данных, наличие средств ускоренного доступа к данным (индексирование, партиционирование, распараллеливание выполнения запросов, кэширование), наличие стандартизированного языка запросов structured query language (далее — SQL), простота интерпретации реляционной модели данных (визуализация в формате табличных структур). Для решения второй задачи популярным решением является Python за счет богатого набора библиотек, простоты понимания кода, широкой поддержки сообщества. Однако, для того чтобы собрать данные для последующего анализа, требуется настроить канал связи между реляционной базой данных и инструментом анализа данных, что влечет за собой несколько проблем, среди которых: затраты на передачу данных (транспортировка гигабайтов информации слишком затратна, особенно с учетом того, что результаты анализа чаще всего приходится транспортировать в ту же реляционную базу данных в отдельные сущности), информационная безопасность (требуется защищать данные,

которые передаются по сети, особенно если речь идет о персональной информации — для таких случаев зачастую дополнительно используют алгоритмы шифрования). Также при обработке большого количества кортежей единого вида и формата реализованные средства доступа к данным в рамках традиционных СУБД могут дать определенные преимущества в контексте переносимости и масштабируемости решения. Следовательно, с течением времени вопрос об интеграции алгоритмов машинного обучения в реляционные СУБД становится все более актуальным.

Об актуальности заявленной проблемы также свидетельствует большое число научных работ, посвященных возможностям современных СУБД для решения задачи интеллектуального анализа данных. Среди наиболее распространенных тематик авторами выделяются вопросы рациональности и эффективности применения СУБД для решения задачи анализа данных в целом [1, 2] и реляционных баз данных в частности, а также исследования, посвященные конкретным СУБД в контексте интеллектуального анализа данных: PostgreSQL [3], Microsoft SQL Server [4], Oracle [5] и вопросам архитектуры эффективных систем с точки зрения обработки и анализа больших объемов данных [6].

2. ОПИСАНИЕ ПРОБЛЕМАТИКИ

В контексте задачи интеграции реляционных СУБД и средств анализа данных продолжают выделять три основных способа [7]:

1. *Слабое связывание.* При слабом связывании инструмент (или система) для анализа данных забирает «сырые» данные из реляционного хранилища, самостоятельно обрабатывает данные с помощью реализованных алгоритмов и далее (при необходимости) возвращает результаты. Подобного связывания тяжело избежать, если источники разнородны (например рукописные оцифрованные отчеты, реляционные и нереляционные базы данных, файлы различных форматов).
2. *Среднее связывание.* Инструмент (или система) для анализа данных по-прежнему самостоятельно выполняет основную часть анализа данных. Однако реляционное хранилище выполняет функции предобработки, подготовки данных, а также пересчета основных статистических характеристик.
3. *Сильное связывание.* Вся обработка данных происходит средствами реляционной СУБД с использованием средств ускорения доступа к данным. Реализация подобного связывания является наиболее трудоемкой, но позволяет значительно сократить накладные расходы на транспортировку данных [8] и повысить производительность при их предобработке и анализе [9].

Если говорить о проблемах, связанных с реализацией последнего из видов связывания, то это, в том числе, ориентированность реляционных СУБД на строки (а не на столбцы), что влечет значительные временные затраты при неэффективной реализации расчета статистических характеристик и алгоритмов машинного обучения. Также актуальной является проблема недостаточности структурного наполнения языка SQL.

В связи с вышеупомянутыми проблемами текущие технические решения по анализу данных средствами СУБД делятся на три основные группы [10]:

1. Отдельные модули и библиотеки, подключаемые или внедряемые в СУБД для решения задач машинного обучения и статистического анализа данных.
2. Расширения языка SQL дополнительными структурами данных и алгоритмами, адаптированными под конкретную СУБД.

3. Индивидуальные реализации алгоритмов машинного обучения на языках SQL и процедурном SQL.

Далее будут подробнее рассмотрены первые две группы, поскольку третья не определяется стандартами, а скорее, отражает индивидуальный взгляд авторов на идею внедрения алгоритмов анализа больших данных в реляционные СУБД.

3. МЕТОДЫ РЕШЕНИЯ ПРОБЛЕМАТИКИ В СОВРЕМЕННЫХ СУБД

Для сравнения возможностей реляционных СУБД в контексте анализа больших данных автором были выбраны следующие средства: Oracle, PostgreSQL, Microsoft SQL Server и MySQL — многофункциональные, распространенные и активно использующиеся в различных технических проектах. По данным DB-Engines, именно эти СУДБ лидируют на российском рынке [11]. Динамика их популярности отображена на рис. 1.

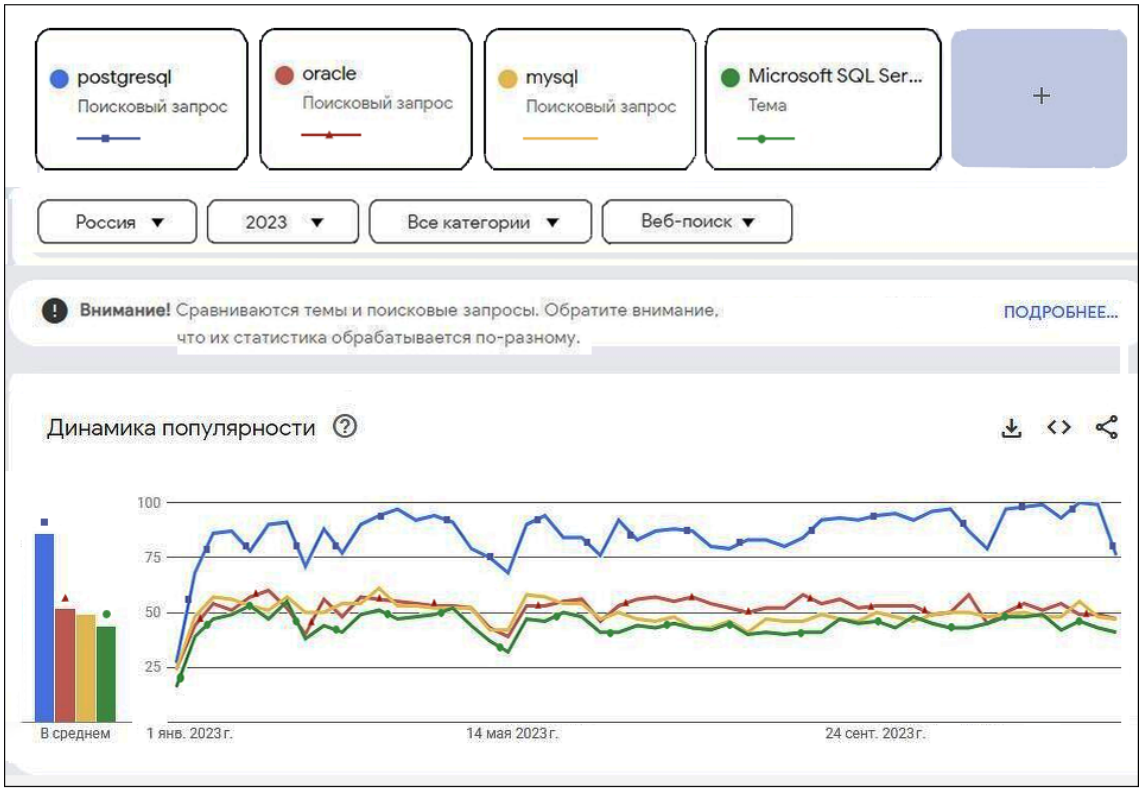

Рис. 1. Динамика популярности современных СУБД в России [11]

Безусловно, в относительно большом количестве разнообразных проектов встречаются и задачи интеллектуального анализа данных. Ниже будет представлен сравнительный анализ ведущих СУБД в части технического обеспечения, предназначенного для решения заявленных задач. Также авторы посчитали нужным включить в этот перечень СУБД DB2. Она не так широко распространена, в первую очередь из-за высокой стоимости, между тем возможности по эффективной обработке данных у этой СУБД не уступают рассматриваемым конкурентам.

3.1. PostgreSQL

Анализ больших данных включает в себя как статистический анализ, так и большой набор алгоритмов обучения (с учителем, без учителя, с подкреплением). Что касается внутреннего статистического анализа, PostgreSQL предоставляет довольно большое количество встроенных агрегатных статистических функций, среди которых:

— коэффициенты корреляции (corr и regr_r2) и ковариации совокупности и выборки (covar_pop и covar_samp соответственно);
— коэффициенты наклона (regr_slope) и смещения (regr_intercept) линейного тренда;
— стандартные отклонения (stddev, stddev_pop и stddev_samp) и дисперсии (variance, var_pop и var_samp).

Для вышеупомянутых функций характерны такие типы входного аргумента, как smallint, integer, bigint, numeric, real и double precision [12].

Для задач интеллектуального анализа данных у СУБД PostgreSQL предусмотрено несколько вариантов внедренных технологий, среди которых в первую очередь следует упомянуть библиотеку MADlib от Apache, а также расширение языка pl/R. Подробный обзор и анализ особенностей этих подходов будет представлен далее.

3.1.1. MADlib

MADlib — библиотека с открытым исходным кодом, которая разработана компанией Apache для внедрения в СУБД PostgreSQL и Greenplum. Архитектура библиотеки представляет следующий набор слоев:

- Интерфейс. Пользователь взаимодействует с библиотекой посредством вызова функции SQL-запросом.
- Слой абстракции на Python, который отвечает за управление алгоритмом, реализованным на pl/python, что приводит к некоторым ограничениям из-за «недоверенности» языка (потенциальные риски и проблемы будут рассмотрены далее).
- Слои основных функций: непосредственно добычи данных и взаимодействия с backend-платформой, реализованных на C++. На 8 сентября 2023 г. вышла уже тринадцатая версия библиотеки — v2.1.0. Она обеспечивает достаточно широкое покрытие возможностями для анализа данных. Библиотека предоставляет различные возможности, в том числе:
  — функции предобработки и подготовки данных (например матричная факторизация — svd или lmf_igd_run, энкодинг категориальных переменных — encode_categorical_variables или стемминг — stem_token() или stem_token_arr());
  — алгоритмы обучения с учителем: деревья решений (tree_train и tree_predict — при этом использован алгоритм CART [13]) и случайный лес (forest_train и forest_predict), knn (с одноименным названием), нейронные сети (а именно, многослойный перцептрон: mlp_classification, mlp_regression для обучения, mlp_predict для предсказания), регрессионные модели (в том числе логистическая регрессия — logregr_train, logregr_predict, мультиномиальная регрессия — multinom, multinom_predict), обобщенные линейные модели — glm и glm_predict, метод опорных векторов — svm_classification, svm_regression для обучения, svm_predict для предсказания;

— алгоритмы обучения без учителя: алгоритм априори (assoc_rules) для поиска ассоциативных правил, кластеризация методом K-means — реализовано несколько разновидностей с учетом различных способов выбора центроид (kmeans_random со случайной инициализацией, kmeanspp — kmeans++ как метод определения центроид, kmeans с явным указанием), методы снижения размерности — PCA (pca_train и pca_sparse_train для разреженных матриц) и PCP (madlib.pca_project и madlib.pca_sparse_project);
— метод ARIMA для анализа временных рядов (arima_train, arima_forecast);
— методы для отбора моделей — кросс-валидация (cross_validation_general), метрики (mean_abs_error, mean_perc_error, r2_score и другие), разделение на тестовые и тренировочные данные (train_test_split) [14].

Из основных преимуществ решения — большое покрытие алгоритмов и функций, использование преимуществ СУБД PostgreSQL и Greenplum и сокращение затрат на транспортировку данных между реляционным хранилищем и отдельным инструментом анализа данных.

Одним из основных недостатков библиотеки MADlib, помимо необходимости в изучении дополнительной документации, установки и интеграции алгоритмов с таблицами в реляционной базе данных, а также недостаточности алгоритмов кластеризации и поиска ассоциативных правил, является использование «недоверенного» языка python в слое абстракции, как уже упоминалось ранее. Согласно документации PostgreSQL, процедурный pl/python доступен исключительно как «ненадежный язык» и не предполагает никаких способов ограничения функциональности. Автор функции на процедурном языке должен самостоятельно решать вопросы, связанные с несанкционированным использованием его функций, поскольку подобные функции обладают правами, аналогичными администратору баз данных. При этом только суперпользователи могут создавать функции на языке pl/python [12]. Если объяснить смысл «недоверенности» простым языком, то при создании и использовании функций и процедур на pl/Python нельзя ограничить права доступа к объектам базы данных на основе какой-либо ролевой модели. Соответственно, в случае с многопользовательскими системами администратору потребуется выдавать пользователям права суперпользователя (что влечет за собой большие риски) — иначе функции и процедуры использовать будет невозможно. Наконец, на официальном сайте не представлен архив для скачивания и установки библиотеки на операционную систему Windows.

3.1.2. pl/R

pl/R — расширение или загружаемый процедурный язык, который позволяет использовать возможности статистического языка R для написания функций и триггеров PostgreSQL. При написании функций используется стандартный синтаксис языка R с незначительными отклонениями (отсутствие закрывающих фигурных скобок или назначения функции). Стоит отметить, что при вызове функции происходит двойное преобразование переменных — сначала переменные приводятся к стандарту языка R, а потом результаты обратно трансформируются в типы, заявленные PostgreSQL. Также, начиная с версии 8.4, PostgreSQL позволяет писать оконные функции на процедурном R. Ниже приведены основные сведения о языке R.

R является интерпретируемым языком, может работать с различными парадигмами программирования, наиболее эффективен в объектно-ориентированном программиро-

вании. В языке R наиболее популярными являются 2 пакета для работы с данными — dplyr и data.table:

— dplyr предоставляет 5 основных функций работы с данными — select (для отбора), filter (для фильтрации), arrange (для сортировки), mutate (для добавления столбцов), summarise (для суммирования части данных);
— data.table использует компактный формат dt[i, j, by], где dt — таблица, i — условие для отбора строк, j — оператор вычислений, by — условие для группировки [15].

R позволяет в том числе:

— использовать бинарную и многоклассовую линейную регрессию (для этого используется метод lm);
— осуществлять классификацию методами knn, деревом решений и случайным лесом;
— использовать алгоритмы снижения размерности (PCA) [16].

Из значимых недостатков, помимо необходимости изучения нового синтаксиса, который отличается от стандартного процедурного SQL, и возможных сложностей в отладке [17], нужно отметить проблемы с «ненадежностью» языка аналогично pl/python, а также невозможность использования процедур, написанных на языке pl/R для создания функций ввода и вывода новых типов данных [18]. Некоторые авторы отмечают проигрыш в скорости обработки данных, особенно при относительно простых задачах [19].

3.2. MySQL

MySQL — популярная реляционная СУБД с богатой историей. В настоящее время поддержка и обслуживание этой СУБД осуществляется компанией Oracle. MySQL также предоставляет возможности для статистического анализа данных средствами СУБД, однако этот набор более скуден: он включает в себя разве что функции стандартного отклонения и дисперсии (std(), stddev(), stddev_pop(), stddev_samp(), var_pop(), var_samp(), variance()) [20].

Из основных решений задачи оптимизации анализа данных отмечают высокопроизводительный ускоритель запросов в памяти MySQL HeatWave, направленный на повышение производительности аналитической и смешанной нагрузки OLTP и OLAP [21].

Из основных преимуществ — отсутствие необходимости транспортировки данных и возможность запуска любым клиентом или приложением, подключенным к базе данных MySQL. Из весомых недостатков — жесткая зависимость от технологии Oracle AutoML в контексте автоматизации обучения моделей машинного обучения.

Соответственно, самостоятельных законченных реализаций аналитических пакетов или модулей для решения задач интеллектуального анализа данных у MySQL на текущий момент в открытых источниках не отмечено.

3.3. Oracle

В рамках СУБД Oracle наибольшую функциональность имеют два решения: Oracle Machine Learning и Oracle Data Miner (как графический интерфейс) совместно с OML for SQL (как набор расширений диалекта SQL).

3.3.1. Oracle Machine Learning

Oracle Machine Learning условно включает в себя OML notebooks, интерфейс для работы с пользователем без кода OML AutoML и средство мониторинга OML Monitoring на автономной базе данных, а также набор расширений языка OML for SQL.

Notebooks предоставляет оболочку для построения моделей машинного обучения в автономной базе Oracle, компонент AutoML обладает графическим интерфейсом для работы с большими данными в контексте данного СУБД, модуль monitoring отвечает за сбор и анализ метрик, относящихся к данным в базе во временном разрезе.

Oracle Machine Learning предлагает большое количество реализованных моделей для интеллектуального анализа данных, среди которых:

— алгоритмы классификации (дерево решений и случайный лес, семантический анализ, градиентный бустинг, наивный Байес, нейронные сети, метод опорных векторов);
— алгоритмы регрессии (градиентный бустинг, нейронные сети, обобщенные линейные модели, метод опорных векторов);
— обработка временных рядов (экспоненциальное сглаживание);
— поиск ассоциативных правил (Apriori);
— кластеризация (k-means, O-cluster, Expectation Maximization);
— методы снижения размерности (например PCA).

При этом настройки алгоритмов реализованы как отдельные объекты в базе данных.

3.3.2. OML for SQL

Наряду с OML for Python и OML for R, OML for SQL предоставляет API для интеллектуального анализа данных с использованием языков SQL и процедурного SQL. Фактически OML4SQL (OML for SQL) состоит из расширений диалекта Oracle SQL и расширяет его возможности в подсчете стоимости запросов, предобработке данных, получении метрик модели [22]. Также он имеет графическую оболочку в виде инструмента Oracle Data Miner, который будет рассмотрен далее.

Oracle Data Miner, в свою очередь, является инструментом-расширением Oracle SQL Developer и обладает возможностями построения, сравнения и генерации моделей машинного обучения, а также предоставляет различные функции (например для предобработки данных — работа с пропусками и выбросами, вычисления метрик и визуализации результатов). Для того чтобы установить Oracle Data Miner, потребуется установить сам Developer (что на данный момент бесплатно), а далее — установить репозиторий Data Miner с помощью скрипта или через графический интерфейс.

Oracle Data Miner использует в своей работе несколько компонентов СУБД Oracle, в том числе:

— Oracle Text — поддерживает анализ текстовых данных средствами СУБД;
— OML for R — позволяет использовать функции, написанные на статистическом языке R;
— Oracle Machine Learning (был представлен ранее) [23].

Фактически OML4SQL реализует интеграцию графического интерфейса в виде Oracle Data Miner с ядром СУБД для построения моделей машинного обучения. Помимо этого,

благодаря технологии Oracle Exadata Smart Scan существует возможность переноса обработки оценки на уровень хранения данных, что значительно повышает производительность оценки данных.

Если говорить подробнее про Smart Scan, то основное преимущество технологии заключается в том, что интенсивные затратные операции выгружаются непосредственно на серверы хранения данных. Благодаря этому считывание и обработка данных происходят параллельно на всех серверах хранения. В свою очередь, на сервер баз данных отправляются только те строки и столбцы, которые непосредственно относятся к запросу [24].

Также СУБД Oracle реализует ряд статистических функций, предназначенных для анализа данных, среди которых:

— функции примерной оценки выражений (APPROX_COUNT, APPROX_MEDIAN, APPROX_PERCENTILE, APPROX_RANK, APPROX_SUM и их разновидности);
— функции работы с битовыми типами — актуально при работе с растровыми изображениями (BIT_TO_NUM, BITMAP_BIT_POSITION, BITMAP_CONSTRUCT_AGG и др.);
— функции работы с кластерами (CLUSTER_DETAILS, CLUSTER_ID, CLISTER_SET и др.);
— функции ковариации и корреляции (COR, CORR_*, COVAR_POP, COVAR_SAMP);
— функции работы с признаками (FEATURE_COMPARE, FEATURE_DETAILS, FEATURE_SET и др.);
— функции определения характеристик выбросов (KURTOTIS_POP, KURTOTIS_SAMP);
— функции предсказания — используются с указанием моделей (PREDICTION_COST, PREDICTION_DETAILS, PREDICTION_SET и др.);
— статистические критерии (STATS_F_TEST, STATS_MW_TEST, STATS_ONE_WAY_ANOVA и др.);
— дисперсии и стандартные отклонения (VAR_POP, VAR_SAMP, VARIANCE, STDDEV_POP, STDDEV, STDDEV_SAMP) [25].

Преимуществ достаточно много: разнообразный набор статистических функций, модули с большим количеством алгоритмов, а также графический интерфейс и расширения языка SQL. Среди существенных недостатков всех модулей и расширений Oracle нужно выделить недоступность на территории РФ и отсутствие открытого кода — зависимость от компании Oracle. Также в качестве особенностей стоит отметить, что Oracle Machine Learning и все его компоненты доступны только для автономной базы данных.

3.4. Microsoft SQL Server

СУБД Microsoft SQL Server также предоставляет пользователям возможность работать со статистическими функциями. Среди прочих, СУБД предлагает следующие функции:

— приблизительного вычисления выражений (APPROX_COUNT_DISTINCT, APPROX_PERCENTILE_DISC. APPROX_PERCENTILE_DISC);
— стандартные отклонения и дисперсии (VAR, VARP, STDEV, STDEVP) [26].

Что касается утилит для работы с большими данными, наиболее распространенным и развитым решением от Microsoft являются службы Analysis Services.

Набор служб Analysis Services — подсистема аналитических данных. Они предоставляют возможности работы с многомерными данными (так называемый OLAP [27], а также интеллектуального анализа больших данных.

Аналитические службы Microsoft реализуют следующие алгоритмы машинного обучения:

— поиск ассоциативных правил;
— деревья решений;

— кластеризация с помощью последовательностей;
— линейная и логистическая регрессии;
— наивный Байес;
— нейронные сети;
— работа с временными рядами — алгоритмы ARIMA, ARTXP [28].

Из преимуществ Analysis Services можно отметить высокую доступность и масштабируемость (балансировка нагрузки NLB, отказоустойчивая кластеризация Windows Server [28]), безопасность с использованием ролей модели Analysis Services, гибкость в выборе источника данных (можно использовать как традиционные реляционные базы данных (Oracle, SQL Server), так и многомерные (Db2 UDB, SAS BW), а среди недостатков — весьма большой набор знаний и навыков, которыми необходимо обладать для использования служб [29]. К тому же продукция Microsoft не доступна на территории РФ.

3.5. DB2

Как правило, под DB2 подразумевают семейство баз данных компании IBM. СУБД IBM DB2 database не так часто используется в проектах из-за достаточно высокой стоимости, она больше подходит для промышленных предприятий с высокими требованиями к безопасности и производительности. Поскольку вопрос анализа больших данных не обошел стороной и эту СУБД, в ней также реализовано большое количество статистических функций. В рамках самой СУБД DB2 реализованы функции:

— ковариации и корреляции (COVARIANCE, CORRELATION);
— регрессионного анализа (REGR_AVGX, REGR_AVGY);
— дисперсии и стандартного отклонения (STDDEV, VARIANCE) [30].

Среди средств для интеллектуального анализа данных, выделяется IBM DB2 BI, состоящее из нескольких компонентов, в том числе Intelligent Miner для интеллектуальной обработки данных. Также важно отметить, что в средстве BI присутствуют как компоненты для взаимодействия с реляционным хранилищем (Warehouse), так и средства поддержки доступа и даже OLAP-модули [31].

В целом, Intelligent Miner адаптирован и активно используется для работы с большими данными, хранящимися в реляционных базах данных (DB2) или файлах.

В этом средстве реализованы среди прочих следующие алгоритмы:

— поиск часто встречающихся наборов;
— кластеризация (с помощью карт Кохонена, BIRCH — balanced iterative reducing and clustering using hierarchies);
— классификация деревом решений, наивным Байесом и логистической регрессией;
— регрессия: полиномиальная, линейная, RBF (radial basis function); а также transform regression — запатентованный алгоритм IBM, совмещающий в себе линейную регрессию и нелинейные преобразования [30];
— работа с временными рядами: ARIMA, экспоненциальное сглаживание и сезонная декомпозиция [30].

Среди преимуществ — более экономная работа с ресурсами, более удобная работа с конфигурациями (в сравнении с PostgreSQL, например) [32]. При этом, это единственное СУБД общего назначения, которое имеет реализации на аппаратно-программном уровне (система IBM i) [33]. Будучи ранее «законодателем мод» реляционных СУБД, сейчас DB2 уже бренд, показатель качества. Из существенных недостатков — большая стоимость

(хоть и есть бесплатный урезанный режим), а также закрытость кода — собственность компании IB2 (как следствие, ограниченная доступность).

4. ВЫВОДЫ

По результатам проведенного обзора и сравнительного анализа возможностей традиционных СУБД в контексте интеллектуального анализа данных можно сделать ряд выводов:

1. Проблема интеграции алгоритмов машинного обучения в современные реляционные СУБД актуальна: это подтверждают теория и практика современных исследований в области систем управления данными.
2. У ведущих СУБД существуют различные технические средства, предназначенные для решения проблемы: модули, дополнительные пакеты, сторонние библиотеки, расширения языка.
3. На сегодняшний день не существует какого-либо «идеального» решения — имеются как проблемы с доступностью (Oracle/SQL Server), так и технические ограничения (PostgreSQL).
4. Алгоритмическая база реализована не полностью — алгоритмы в различных решениях не покрывают весь спектр возможностей машинного обучения. Это является проблемой, поскольку для решения различных задач оптимизированы разные алгоритмы.
5. С учетом доступности и широкой поддержки сообщества на территории РФ остается открытой ниша развития интеллектуального анализа данных средствами СУБД PostgreSQL. Поэтому научные исследования в этой области несомненно представляют большой интерес.

Список литературы

1. *Худяков В.Б.* Использование СУБД в проектах машинного обучения и анализа данных // Вестник науки. 2023. № 7. С. 278–295.
2. *Наумов Р. К., Самылкин М. С., Копейкин М. В.* Способы интеллектуального анализа данных средствами СУБД // Научный результат. Информационные технологии. 2021. № 2. С. 32–40.
3. *Постойко А. Ю.* Интеграция нейронных сетей в СУБД PostgreSQL // Актуальные проблемы авиации и космонавтики: сборник материалов VIII Международной научно-практической конференции. Т. 2. Красноярск, 11–15 апреля 2022. С. 167–169.
4. *Аверьянова Е. В.* Средства интеллектуального анализа данных в Microsoft SQL Server // Экономика и социум,. 2016. № 11. С. 324–326.
5. *Соболева А. Д., Сабинин О. Ю.* Разработка метода композиции алгоритмов машинного обучения для решения задачи прогнозирования на примере технологии Oracle Data Mining // Theoretical & Applied science. 2018. № 3. С. 147–154.
6. *Часовских В. П., Кох О. С.* СУБД Greenplum для Big Data и машинного обучения // BI-технологии и корпоративные информационные системы в оптимизации бизнес-процессов цифровой экономики: материалы X Международной научно-практической очно-заочной конференции. Екатеринбург, 2 декабря 2022. С. 116–120.
7. *Saragawi S., Thomas S., Agrawal R.* Integration Association Rule Mining with Relational Database Systems: Alternatives and Implications // SIGMOD: Proceedings ACM SIGMOD International Conference on Management of Data. Seattle, June 2–4, 1998. P. 343–354.

8. *Миниахметов Р. М., Цымблер М. Л.* Интеграция алгоритма кластеризации Fuzzy C-means в PostgreSQL // Вычислительные методы и программирование. 2012. № 2. С. 46–52.
9. *Han J., Kamber M.* Data Mining: Concepts and Techniques. Amsterdam: Morgan Kaufmann, 2006.
10. *Цымблер М. Л.* Обзор методов интеграции интеллектуального анализа данных в СУБД // Вестник Южно-Уральского Государственного Университета. 2019. № 2. С. 32–62.
11. PostgreSQL возглавила мировой рейтинг роста популярности СУБД и стала абсолютным лидером среди популярных СУБД в России [Электронный ресурс]. URL: https://www.cnews.ru/news/line/2024-01-09_postgresql_vozglavila_mirovoj (дата обращения: 10.03.2024).
12. Документация СУБД PostgreSQL [Электронный ресурс]. URL: https://postgrespro.ru/docs/postgresql/16/index (дата обращения 17.03.2024).
13. *Breiman L., Friedman J., Olshen R. A., and others.* Classification and Regression Trees, Monterrey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
14. Документация библиотеки Apache MADlib [Электронный ресурс]. URL: https://madlib.apache.org/docs/latest/index.html (дата обращения: 25.03.2024).
15. *Гибадуллина Д. А.* Язык программирования R для статистической обработки данных [Электронный ресурс]. URL: https://habr.com/ru/articles/781086/ (дата обращения: 02.04.2024).
16. Machine Learning with R [Электронный ресурс]. URL: https://github.com/PacktPublishing/Machine-Learning-with-R-Third-Editin (дата обращения: 10.04.2024).
17. *Conway J.* Easy Statistical Analysis in PostgreSQL with PL/R: presentation // PgDay'15 Russia: the second official Russian Conference. Saint-Petersburg, July 16, 2015.
18. Документация расширения pl/R [Электронный ресурс]. URL: https://github.com/postgres-plr/plr/blob/master/userguide.md (дата обращения: 15.04.2024).
19. Статистический анализ в PostgreSQL с помощью PL/R [электронный ресурс]. URL: https://habr.com/ru/articles/275487/ (дата обращения: 20.04.2024).
20. Документация СУБД MySQL [Электронный ресурс]. URL: https://dev.mysql.com/doc/refman/8.0/en/aggregate-functions.html (дата обращения: 23.04.2024).
21. Документация HeatWave [Ээлектронный ресурс]. URL: https://dev.mysql.com/doc/heatwave/en/mys-hw-introduction.html (дата обращения: 26.04.2024).
22. Документация Oracle Machine Learning [Электронный ресурс]. URL: https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/21/mlsql/oracle-machine-learning-sql.html#GUID-7D00AFBD-EDED-418C-81FB-576A83CA9536 (дата обращения: 29.04.2024).
23. Документация Oracle Data Miner [Электронный ресурс]. URL: https://docs.oracle.com/en/database/oracle/sql-developer/23.1/dmrig/oracle-data-miner-installation-and-administration-guide.pdf (дата обращения: 02.05.2024).
24. Описание Oracle SmartScan [Электронный ресурс]. URL: https://www.oracle.com/database/technologies/exadata/software/smartscan/ (дата обращения: 05.05.2024).
25. Документация Oracle [Электронный ресурс]. URL: https://docs.oracle.com/en/database/oracle/oracle-database/21/sqlrf/Functions.html#GUID-D079EFD3-C683-441F-977E-2C9503089982 (дата обращения: 10.05.2024).
26. Справочник по Transact-SQL [Электронный ресурс]. URL: https://learn.microsoft.com/ru-ru/sql/t-sql/functions/aggregate-functions-transact-sql?view=sql-server-ver16 (дата обращения: 15.05.2024).
27. *Codd E., Codd S., Salley C.* Providing OLAP to User-Analysts: An IT Mandate. Codd & Associates, 1993.
28. Документация Analysis Services (раздел алгоритмов интеллектуального анализа данных) [Электронный ресурс]. URL: https://learn.microsoft.com/en-us/analysis-services (дата обращения: 20.05.2024).
29. *Panchal R.* 21+ Pros and Cons of Azure Analysis Services [Электронный ресурс]. URL: https://thenextfind.com/pros-cons-of-azure-analysis-services/ (дата обращения: 28.05.2024).
30. Документация IBM DB2 [Электронный ресурс]. URL: https://www.ibm.com/docs/en/db2/9.7?topic=functions-user-defined (дата обращения: 01.06.2024).
31. Анализ данных с целью поддержки принятия решений (IBM DB2 Business Intelligence) [Электронный ресурс]. URL: https://intuit.ru/studies/courses/85/85/lecture/28289?page=4 (дата обращения: 05.06.2024).

32. IBM DB2 для 1С: Предприятие [Электронный ресурс]. URL: https://www.handybackup.ru/1c-db2.shtml (дата обращения: 20.06.2024).
33. *Драч В.* Сравнение современных СУБД [Электронный ресурс]. URL: https://drach.pro/blog/hi-tech/item/145-db-comparison (дата обращения: 25.06.2024).

Поступила в редакцию 13.06.2024, окончательный вариант — 27.06.2024.

Графеева Наталья Генриховна, канд. физ.-мат. наук, доцент, доцент, университет ИТМО, nggrafeeva@corp.ifmo.ru

Назаров Артем Александрович, студент магистратуры, университет ИТМО, ✉ artem.a.nazarov@yandex.ru

Computer tools in education, 2024
№ 2: 58–71
http://cte.eltech.ru
doi:10.32603/2071-2340-2024-2-58-71

Implementation of Machine Learning Algorithms by Means of Relational Database Management Systems

Grafeeva N. G.[1], Cand. Sc., Associate Professor, nggrafeeva@corp.ifmo.ru
Nazarov A. N.[1], Student, ✉ artem.a.nazarov@yandex.ru

[1]ITMO University, 49 Kronverksky, bldg. A, 197101, Saint Petersburg, Russia

Abstract

The article discusses the problems of integrating machine learning algorithms into relational database management systems. The author conducted a review and comparative analysis of the current technical capabilities of relational database management systems like Oracle, PostgreSQL, SQL Server, DB2 and MySQL, which were adapted to data mining. Based on the obtained results, conclusions about the level of readiness of modern database management systems to solve the problem of data analysis have been drawn.

Keywords: *relational database management systems, data mining, technical solutions.*

Citation: N. G. Grafeeva and A. N. Nazarov, “Implementation of Machine Learning Algorithms by Means of Relational Database Management Systems,” *Computer tools in education*, no. 2, pp. 58–71, 2024 (in Russian); doi:10.32603/2071-2340-2024-2-58-71

References

1. V. Khudiakov, “Using DBMS in machine learning and data analysis projects,” *Vestnik Nauki*, no. 7,pp. 278–295, 2023 (in Russian).
2. R. K. Naumov, M. S. Samylkin, and M. V. Kopeikin, “Data Mining Methods Using DBMS Tools,” *Research result. Information technologies*, vol. 6, no. 2, pp. 32–40, 2021 (in Russian); doi:10.18413/2518-1092-2021-6-2-0-5
3. A. Y. Postoyko,“Neural networks integration into POSTGRESQL DBMS,”in *Proc. of Aktual’nye problemy aviatsii i kosmonavtiki, Krasnoyarsk, Russia, Apr. 11-15, 2022*, vol. 2, pp. 167–169, 2022 (in Russian).
4. E. V. Averyanova, “Means data mining Microsoft SQL SERVER,” *Ekonomika i sotsium*, no. 11, pp. 324–326, 2016 (in Russian).

5. A. D. Soboleva and O. Y. Sabinin, “Ensemble learning method development for solving the prediction problem on the example of oracle data mining technology,” *Theoretical & Applied Science*, vol. 59, no. 03, pp. 147–154, 2018 (in Russian); doi:10.15863/tas.2018.03.59.24
6. V. P. Chasovskikh and O. S. Kokh, “Greenplum DBMS for big data and machine learning,”in *Proc. of BI technologies and corporate information systems in optimizing business processes in the digital economy, Ekaterinburg, Russia, Dec. 2, 2022*, pp. 116–120, 2022 (in Russian).
7. S. Saragawi, S. Thomas, and R. Agrawal, “Integration Association Rule Mining with Relational Database Systems: Alternatives and Implications,” in *Proc. of ACM SIGMOD Int. Conf. on Management of Data, Seattle, WA, US, June 2–4, 1998*, pp. 343–354, 1998.
8. R. M. Miniakhmetov and M. L. Tsymbler, “Integration of Fuzzy c-Means Clustering algorithm with PostgreSQL database management system,” *Numerical Methods and Programming*, vol. 13, no. 2, pp. 46–52, 2012.
9. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Amsterdam: Morgan Kaufmann, 2006.
10. M. L. Tsymbler, “Overview of Methods for Integrating Data Mining into DBMS,” *Bulletin of the South Ural State University. Series “Computational Mathematics and Software Engineering”*, vol. 8, no. 2, 2019 (in Russian); doi:10.14529/cmse190203
11. “PostgreSQL topped the global ranking of DBMS popularity growth and became the absolute leader among popular DBMSs in Russia,” in *www.cnews.ru*, 2024 (in Russian). [Online]. Available: https://www.cnews.ru/news/line/2024-01-09_postgresql_vozglavila_mirovoj
12. The PostgreSQL Global Development Group, “PostgreSQL 16.3 Documentation,” in *postgrespro.com*, 2024. [Online]. Available: https://postgrespro.com/docs/postgresql/16/index
13. L. Breiman et al., *Classification and Regression Trees*, Monterrey, CA, US: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
14. “User Documentation for Apache MADlib,” in *madlib.apache.org*, 2023. [Online]. Available: https://madlib.apache.org/docs/latest/index.html.
15. D. A. Gibadullina,“R programming language for statistical data processing,” in *habr.com*, 2023. [Online] (in Russian). Available: https://habr.com/ru/articles/781086/
16. PacktPublishing, “Machine Learning with R,” in *github.com*, 2022. [Online]. Available: https://github.com/PacktPublishing/Machine-Learning-with-R-Third-Edition
17. J. Conway, “Easy Statistical Analysis in PostgreSQL with PL/R,” in *PgDay’15 Russia: the second official Russian Conference. Saint-Petersburg, July 16, 2015*, [Online Presentation], 2015. Available: https://joeconway.com/presentations/plr-DWDC-2015.05.pdf
18. J. Conway, “PL/R User’s Guide - R Procedural Language,” in *github.com*, 2023. [Online]. Available: https://github.com/postgres-plr/plr/blob/master/userguide.md
19. R. Druzyagin, “Statistical analysis in PostgreSQL using PL/R,” in *habr.com*, 2016 (in Russian). [Online]. Available: https://habr.com/ru/articles/275487/
20. Oracle Corp., “MySQL 8.0 Reference Manual,” in *dev.mysql.com*, 2024. [Online]. Available: https://dev.mysql.com/doc/refman/8.0/en/aggregate-functions.html
21. Oracle Corp., “HeatWave User Guide,” in *dev.mysql.com*, 2024. [Online]. Available: https://dev.mysql.com/doc/heatwave/en/mys-hw-introduction.html
22. Oracle Corp., “Machine Learning for SQL Use Cases,” in *docs.oracle.com*, 2024. [Online]. Available: https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/21/mlsql/oracle-machine-learning-sql.html#GUID-7D00AFBD-EDED-418C-81FB-576A83CA9536
23. H. Moitreyee et al., “Oracle Data Miner. Installation and Administration Guide,” in *docs.oracle.com*, 2024. [Online]. Available: https://docs.oracle.com/en/database/oracle/sql-developer/23.1/dmrig/oracle-data-miner-installation-and-administration-guide.pdf
24. Oracle Corp.,“Oracle Exadata Database Machine Smart Scan,” in *docs.oracle.com*, 2024. [Online]. Available: https://www.oracle.com/database/technologies/exadata/software/smartscan/
25. Oracle Corp., “SQL Language Reference. 7 Functions,” *docs.oracle.com*, 2024. [Online]. Available: https://docs.oracle.com/en/database/oracle/oracle-database/21/sqlrf/Functions.html#GUID-D079EFD3-C683-441F-977E-2C9503089982
26. Microsoft Corp., “Aggregate Functions (Transact-SQL),” in *learn.microsoft.com*, 2023. [Online]. Available: https://learn.microsoft.com/en-us/sql/t-sql/functions/aggregate-functions-transact-sql?view=sql-server-ver16
27. E. Codd, S. Codd, and C. Salley, *Providing OLAP to User-Analysts: An IT Mandate*, US: Codd & Associates, 1993.
28. Microsoft Corp., “Analysis Services documentation,” in *learn.microsoft.com*, 2024. [Online]. Available: https://learn.microsoft.com/en-us/analysis-services
29. R. Panchal,“21+ Pros and Cons of Azure Analysis Services,” in *thenextfind.com*, 2024. [Online]. Available: https://thenextfind.com/pros-cons-of-azure-analysis-services/
30. IBM Corp., “DB2 Version 9.7 for Linux, UNIX, and Windows,” in *www.ibm.com*, 2024. [Online]. Available: https:

//www.ibm.com/docs/en/db2/9.7?topic=functions-user-defined
31. V. Varfolomeev, “Data Analysis for Decision Support (IBM DB2 Business Intelligence),” in *intuit.ru*, 2011 (in Russian). [Online]. Available: https://intuit.ru/studies/courses/85/85/lecture/28289?page=4
32. Novosoft LLC, “IBM DB2 for 1C: Enterprise,” in *www.handybackup.ru*, 2024. [Online]. Available: https://www.handybackup.ru/1c-db2.shtml
33. V. Drach, “Comparison of modern DBMS,” in *drach.pro*, 2017 (in Russian). [Online]. Available: https://drach.pro/blog/hi-tech/item/145-db-comparison

Received 13-06-2024, the final version — 27-06-2024.

Natalia Grafeeva, Cand. of Sciences (Phys.-Math.), Associate Professor, ITMO University, nggrafeeva@corp.ifmo.ru

Artem Nazarov, Master’s Degree student, ITMO University, ✉ artem.a.nazarov@yandex.ru