Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ_____

КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ_____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Оценка количества разных значений в столбце таблицы реляционной базы данных методом машинного обучения

Студент __ИУ5-32М_____ _____ **Лу Сяои**_____
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Ю.Е. Гапанюк**_____
(Подпись, дата) (И.О.Фамилия)

2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_04_» ____сентября_____ 2023 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме _____
_____

Студент группы _ИУ5-32М__

_____Лу Сяои_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
_____ИССЛЕДОВАТЕЛЬСКАЯ_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____

График выполнения НИР: 25% к ____ нед., 50% к ____ нед., 75% к ___ нед., 100% к ____ нед.

Техническое задание _____
Оценка количества разных значений в столбце таблицы методом машинного обучения_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __19___ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания «_04_» ____сентября_____ 2023 г.

Руководитель НИР _____ _____Ю.Е. Гапанюк__
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____Лу Сяои_____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение 4

Глава 1:Исследовательская база для оценки количества различных значений (NDV) 5

1.1 Сочетание машинного обучения и баз данных 5

1.2 Концепция оценки NDV/кардинальности 5

1.3 Метод оценки NDV/КАРДИНАЛЬНОСТЬ 7

1.3.1 Традиционные методы оценки 7

1.3.2 Оценка кардинальности на основе машинного обучения 8

1.3.3 Метод выбора 9

Глава 2: Преимущества и целесообразность метода оценки 11

2.1 Преимущество оценки максимального правдоподобия (MLE), основанной на выборках 11

2.1.1 Метод отбора образцов 11

2.1.2 Оценка максимального правдоподобия 12

2.2 Генерация обучающих данных 13

2.3 MLE 13

2.4 Существующие оценщики 14

2.5 Структура модели 15

Заключение 17

Список использованной литературы 18

Введение

Оценка количества отдельных значений(NDV) в столбце полезна для реш ения многих задач в системах баз данных, таких как выбор методов сжатия стол бцов в СУБД,основанных на столбцах, профилирование данных.

Целью данной работы является разработка метода оценки количества отд ельных значений(NDV) в столбце таблицы базы данных, при этом основное вни мание уделяется тому, как получить точные оценки NDV из случайной выборки. Предполагаемый метод заключается в формулировке проблемы как оценки мак симального правдоподобия(MLE).

В данном отчёте представлена история вопроса, проведен обзор других не давних исследований по оценке NDV/кардинальности в мире, а также сравнени е характеристик различных методов оценки кардинальности. Затем анализирую тся преимущества метода оценки. MLE-оценщик не зависит от рабочей нагрузк и, поэтому выученную модель можно использовать на невидимых рабочих нагр узках. Наконец, обсуждается осуществимость метода оценки: генерация обуча ющих данных, формулировка задачи в виде решения оценки максимального пр авдоподобия.

Глава 1:Исследовательская база для оценки количества различн ых значений (NDV)

1.1 Сочетание машинного обучения и баз данных

Традиционная технология баз данных часто зависит от алгоритмов вдохновения или ручного вмешательства, таких как настройка параметров базы данных, устранение неполадок, рекомендация индексов и т.д. Однако в эпоху больших данных появляется все больше и больше экземпляров баз данных, все более сложные сцены и большие объемы данных, так что традиционной технологии баз данных трудно удовлетворить потребности больших данных. Например, облачная база данных имеет миллион экземпляров базы данных. Сценарии применения каждого экземпляра и уровень использования пользователями могут быть очень разными. Трудно получить удовлетворительные результаты, используя традиционные алгоритмы вдохновения, а ручное вмешательство затруднительно для управления таким количеством экземпляров. Машинное обучение находит наилучшее дизайнерское решение на основе исторических данных и поведения пользователей и может освободить от множества сложной ручной работы. Именно поэтому технология машинного обучения широко используется во многих областях исследований и производства.

Машинное обучение также открывает новые возможности для технологии оптимизации баз данных. Традиционные модели машинного обучения, такие как линейные датчики, случайные леса, машины опорных векторов и интегрированное обучение, могут накапливать "опыт" в исторических данных и улучшать способность модели решать сложные задачи[1].

1.2 Концепция оценки NDV/кардинальности

В последние годы, с быстрым развитием искусственного интеллекта, технология пересечения баз данных и искусственного интеллекта постепенно становится одним из важных направлений в области исследования баз данных. Оценка числа различимых значений (NDV) в базе данных является одной из наиболее фундаментальных проблем в области баз данных, и оценка значения

NDV может быть применена для улучшения оптимизации запросов, сжатия данных и других проблем.

Оценка количества неповторяющихся значений NDV используется при оценке скорости отбора. Точность NDV напрямую влияет на точность скорости выбора Selectivity, которая, в свою очередь, влияет на точность размера промежуточного результата, на то, является ли оценка затрат обоснованной, и на то, является ли план выполнения оптимальным.

Оптимизатор базы данных (Query Optimizer) является одним из наиболее важных модулей базы данных, который определяет выбор плана запроса (Query Plan) для SQL-запросов и напрямую влияет или даже определяет скорость выполнения запроса.

Проще говоря, цель Cardinality Estimation - предсказать количество строк, которые запрос, скорее всего, вернет без выполнения SQL-запроса.

Например, простой SQL-запрос:

SELECT * FROM DB WHERE A = 1;

Оценка кардинальности является результатом оценки следующего запроса:

SELECT COUNT(*) FROM DB WHERE A = 1;

Оптимизатор запросов использует результаты оценки кардинальности для создания наилучшего плана запроса. При более точной оценке оптимизатор запросов обычно способен генерировать лучшие планы запросов.

Рисунок 1 Принципы работы оптимизатора запросов

Современные оптимизаторы баз данных в значительной степени полагаются на свои внутренние системы оценки стоимости, и наиболее важной основой для оценки стоимости является база оператора запроса, т.е. количество строк результата, оставшихся после того, как данные были отфильтрованы условиями запроса внутри оператора.

Таким образом, техника оценки кардинальности является наиболее важной техникой, которая влияет на производительность плана выполнения, генерируемого оптимизатором.

Техника оценки кардинальности исследовалась и разрабатывалась в течение десятилетий в академических и промышленных кругах, но из-за необходимости баланса между точностью и эффективностью, она по-прежнему является одной из самых сложных тем для решения в базах данных и известна как "ахиллесова пята" оптимизаторов[2].

1.3 Метод оценки NDV/КАРДИНАЛЬНОСТЬ

<table>
<tr><th>Категория</th><th colspan="2">Метод</th></tr>
<tr><td rowspan="3">Традиционный метод</td><td rowspan="2">на основе синопсиса</td><td>Гистограмма</td></tr>
<tr><td>Скетчинг</td></tr>
<tr><td colspan="2">Выборка</td></tr>
<tr><td rowspan="5">Метод, основанный на обучении</td><td rowspan="2">Ориентированный на запросы</td><td>Регрессия</td></tr>
<tr><td>Нейронная сеть</td></tr>
<tr><td rowspan="3">Ориентированные на данные</td><td>Метод ядерной плотности</td></tr>
<tr><td>Метод авторегрессии</td></tr>
<tr><td>Продуктовые сети Sum</td></tr>
</table>

Таблица 1 Похожие работы по методам оценки кардинальности

1.3.1 Традиционные методы оценки

Существующие традиционные методы оценки кардинальности баз данных делятся на три типа методов: гистограммный, эскизный и выборочный[1].

Подход на основе гистограммы нацелен на реальные базы данных, где распределение значений каждого столбца не является равномерным, и

гистограмма поддерживается для сохранения информации каждого диапазона данных, чтобы соответствовать истинному распределению данных.

Оценка кардинальности на основе эскизов опирается на предварительно собранную статистическую информацию из таблицы. Общим подходом к оценке кардинальности ансамбля является использование хэш-функции для отображения данных на битовую карту, а затем выполнение оценки кардинальности на битовой карте, и исследование и улучшение таких алгоритмов включает пункт "MinCount".

Для оценки кард инальности запросов с объединением нескольких таблиц используются эмпирические формулы, которые подвержены большой погрешности оценки кардинальности.

С уменьшением стоимости памяти и появлением баз данных in-memory широко изучаются методы оценки кардинальности на основе выборки. Основной подход заключается в применении запроса к небольшому набору данных, отобранных из исходных данных, для получения оценки общего распределения, и этот подход был использован в in-memory database Hyper.

1.3.2 Оценка кардинальности на основе машинного обучения

Традиционные статистические методы оценки кардинальности страдают от больших ошибок, методы, основанные на выборке, страдают от эффективного распада выборки при оценке многотабличных запросов, а методы выборки, основанные на индексах, сильно зависят от структуры индекса. Последние работы показали, что эффективные, точные и надежные методы оценки кардинальности могут быть достигнуты с помощью машинного обучения[1].

Оценка кардинальности с помощью алгоритмов машинного обучения может быть в целом классифицирована на две категории, одна из которых - моделирование с использованием только самого запроса. Этот подход извлекает характеристики только из самого оператора запроса, не требуя от оценщика кардинальности доступа к таблице данных или точного знания того, как выполняется запрос. Однако точность и стабильность не гарантируются.

Другой тип моделирования процесса выполнения опроса основан на оптимизаторе опроса, который поддерживает оценку кардинальности на уровне плана выполнения. Кроме того, он также может изучать стоимость планов выполнения напрямую, без участия оптимизатора, что избавляет от необходимости настраивать параметры модели стоимости и делает модель оценки кардинальности более устойчивой в конкретной среде с возможностью автоматической миграции в различные среды.

Существующие методы оценки NDV можно в целом разделить на две категории: 1) методы на основе сканирования, которые сканируют все данные и сохраняют эскиз для приблизительной оценки NDV; и 2) методы на основе выборки, которые оценивают NDV, используя выборочные данные, а не доступ ко всему хранилищу данных. Методы, основанные на сканировании, обеспечивают более низкую ошибку аппроксимации за счет более высокого ввода-вывода и большего времени. Оценка на основе выборки предпочтительна в приложениях с большим объемом данных и допустимым ограничением погрешности из-за более высокой масштабируемости.

По мере увеличения объема данных методы оценки NDV на основе сканирования с трудом справляются с большими данными, поэтому все большее значение приобретают методы оценки NDV на основе выборки.

1.3.3 Метод выбора

Существующие методы оценки NDV можно в целом разделить на две категории[3]:

1) методы на основе сканирования, которые сканируют все данные и сохраняют эскиз для приблизительной оценки NDV;

2) методы на основе выборки, которые оценивают NDV, используя выборочные данные, а не доступ ко всему хранилищу данных. Методы, основанные на сканировании, обеспечивают более низкую ошибку аппроксимации за счет более высокого ввода-вывода и большего времени. Оценка на основе выборки предпочтительна в приложениях с большим

объемом данных и допустимым ограничением погрешности из-за более высокой масштабируемости.

По мере увеличения объема данных методы оценки NDV на основе сканирования с трудом справляются с большими данными, поэтому все большее значение приобретают методы оценки NDV на основе выборки.

Глава 2: Преимущества и целесообразность метода оценки

2.1 Преимущество оценки максимального правдоподобия (MLE), основанной на выборках

2.1.1 Метод отбора образцов

Существует долгая история работы по изучению моделей для оценки кардинальности или селективности запросов. В целом, их можно разделить на два типа [4]. Оценка базы запросов включает два типа: ориентированный на запросы и ориентированный на план выполнения.

Первый тип использует запросы в качестве основной характеристики, а обучающие данные поступают из записей выполнения запросов. Например, в [5] запросы используются в качестве характеристик, а для обучения селективности предикатов многомерного диапазона применяются ансамбли на основе деревьев. Такие подходы обычно требуют выполнения большого количества запросов для получения достаточного количества обучающих данных [4]. Подходы, ориентированные на запросы, используют сверточные нейронные сети для изучения взаимосвязей между данными таблицы, условиями запроса и условиями присоединения, однако они требуют длительного обучения и плохо обобщаются в различных сценариях.

Второй тип (например, [6] и [7]) строит модель для аппроксимации совместного распределения всех атрибутов в таблице. Подходы, ориентированные на план исполнения, выполняют каскадную оценку затрат на уровне физического оператора, чтобы в некоторой степени улучшить адаптацию к различным запросам. Такие методы необходимо переобучать в случае обновления данных или схемы.

В динамических средах с неизвестными наборами данных и рабочими нагрузками оба типа методов требуют значительных усилий для доступа к новым наборам данных и переобучения [5].

Существующие методы оценки на основе выборки обычно опираются на эвристику или предположения, которые не отличаются надежностью при

использовании различных наборов данных, поскольку предположения о данных могут быть легко нарушены.

Хотя в этой работе также изучается модель для оценки NDV/кардинальности, подход значительно отличается от описанного выше. Подход, используемый в данной работе, основан на выборке и использует функцию выборки (взятую из таблицы или результатов запроса), а не запрос.Что еще более важно, этот метод не зависит от рабочей нагрузки и может быть применен к любой динамической рабочей нагрузке.

2.1.2 Оценка максимального правдоподобия

Существующие методы оценки на основе выборки обычно опираются на эвристику или предположения, которые не отличаются надежностью при использовании различных наборов данных, поскольку предположения о данных могут быть легко нарушены.

Мы можем сформулировать оценку NDV на основе выборки, используя метод, основанный на принципах принципа для оценки неизвестных параметров, - оценку максимального правдоподобия (MLE) [8]. Оценщик MLE не зависит от нагрузки: он выводится (аналитически) до того, как мы увидим реальную нагрузку. Он решает задачу оптимизации, которая максимизирует вероятность наблюдения определенной случайной выборки, и дает оценку NDV с желаемыми свойствами, такими как согласованность и эффективность.

Оценка NDV по случайным выборкам может быть сформулирована как задача оценки максимального правдоподобия (MLE), которая обычно используется для оценки неизвестных параметров статистических моделей с желательными свойствами, такими как согласованность и эффективность.

В данной работе предлагается и исследуется вопрос: возможно ли обучить модель машинного обучения, не зависящую от рабочей нагрузки, для аппроксимации принципиальных статистических оценок, таких как MLE-оценки, с обучающим набором, синтетически сгенерированным из обучающего распределения, откалиброванного на основе свойств нашей задачи оценки, так, чтобы обученная модель могла использоваться на невидимых рабочих

нагрузках. Такая эффективная оценка необходима для задач, в которых сканирование данных даже один раз является слишком сложной задачей.

2.2 Генерация обучающих данных

Можно использовать модель обучения на синтетических данных без каких-либо реальных данных.

Оптимальный выбор размера обучающих данных для разных колонок может быть разным. Целью данной статьи является обучение оценщика, не зависящего от рабочей нагрузки, и настройка размера обучающих данных для каждой разной рабочей нагрузки может быть слишком дорогой (если не невозможной)[9].

2.3 MLE

Оцененное значение (или решение задачи MLE) максимизирует вероятность наблюдаемых данных, полученных на основе этой модели.

Мы предполагаем, что столбец C с профилем F взят из некоторого предварительного распределения вероятности. Затем из C берется равномерно случайная выборка S с профилем f.

Формула на основе MLE для оценки NDV может быть получена на основе наблюдаемого профиля в выборке S, т.е. профиля выборки f , и наблюдаемого размера популяции N (или, эквивалентно, частоты выборки r = n/N). Мы оцениваем D как тот, который максимизирует вероятность наблюдения f и N[9].

$$\mathrm{D}^{\mathrm{MLE}} = \arg\max_{\mathrm{D}} \mathbb{P}(\mathrm{f}, \mathrm{N}|\mathrm{D}) = \arg\max_{\mathrm{D}} \mathbb{P}(\mathrm{f}, \mathrm{N}|\mathrm{F})\mathbb{P}(\mathrm{F}|\mathrm{D})$$

Определите $\mathcal{F}(D,N) = \left\{F \middle| \sum_{j>0} F_j = D \; and \sum_{j>0} j \cdot F_j = N \right\}$ все выполнимые конфигурации профиля с NDV равным D и размером популяции равным N. Для $\mathrm{F} \in \mathcal{F}(D,N)$ у нас есть $\mathbb{P}(\mathrm{f}, \mathrm{N}|\mathrm{F}) = \mathbb{P}(\mathrm{f}|\mathrm{F})$; для $\mathrm{F} \notin \mathcal{F}(D,N)$ имеем $\mathbb{P}(\mathrm{f}, \mathrm{N}|\mathrm{F}) = 0$ и $\mathbb{P}(\mathrm{F}|\mathrm{D}) = 0$. Приведенная выше формулировка может быть переписана как:

$$\mathrm{D}^{\mathrm{MLE}} = \arg\max_{\mathrm{D}} \sum_{\mathrm{F} \in \mathcal{F}(\mathrm{D},\mathrm{N})} \mathbb{P}(\mathrm{f}|\mathrm{F})\, \mathbb{P}(\mathrm{F}|\mathrm{D})$$

Оценщик $\mathrm{D}^{\mathrm{MLE}}$ должен использоваться на неизвестных столбцах; поэтому для решения вышеприведенной задачи оптимизации разумно предположить, что предварительное распределение F является равномерным, в том смысле,

что каждый возможный профиль в $\mathcal{F}(D,N)$ появляется с равной вероятностью, т.е. $\mathbb{P}(\mathrm{F}|\mathrm{D}) = \frac{1}{|\mathcal{F}(\mathrm{D},\mathrm{N})|}$ для каждого $\mathrm{F} \in \mathcal{F}(\mathrm{D},\mathrm{N})$. При этом предположении мы хотим решить следующую задачу для $\mathrm{D}^{\mathrm{MLE}}$:

$$\mathrm{D}^{\mathrm{MLE}} = \arg\max_{\mathrm{D}} \frac{1}{|\mathcal{F}(\mathrm{D},\mathrm{N})|} \sum_{\mathrm{F}\in\mathcal{F}(\mathrm{D},\mathrm{N})} \mathbb{P}(\mathrm{f}|\mathrm{F}) \qquad (1)$$

Мы также можем интерпретировать формулу (1) на основе MLE следующим образом. После наблюдения профиля выборки f и размера популяции N, мы оцениваем D как $\mathrm{D}^{\mathrm{MLE}}$, которое максимизирует среднюю вероятность генерации f из выполнимого профиля $\mathrm{F} \in \mathcal{F}(\mathrm{D},\mathrm{N})$. Решение (1), однако, трудно даже приблизительно, и поэтому эта формулировка пока не применяется для оценки NDV.

Естественный вопрос заключается в том, выгодно ли использовать выборку S, вместо профиля выборки f, в качестве наблюдаемых данных для получения MLE и в качестве характеристик в нашей системе обучения. На самом деле, большинство существующих оценок (как мы покажем далее) также используют профиль выборки вместо выборки для оценки NDV.

2.4 Существующие оценщики

Существует длинный ряд работ по оценке NDV по случайным выборкам. Мы рассмотрим некоторые из них следующим образом.

•Задача, связанная с задачей в (1), - это профильная оценка максимального правдоподобия (PML) [10, 11, 12], которая выбирает F, максимизирующую вероятность наблюдения f из случайно взятой S.

Определите:

$$\mathrm{F}^{\mathrm{PML}} = \arg\max_{\mathrm{F}} \mathbb{P}(\mathrm{f}|\mathrm{F}) \text{ and } \mathrm{D}^{\mathrm{PML}} = \sum_{\mathrm{j}>0} \mathrm{F}_{\mathrm{j}}^{\mathrm{PML}} \qquad (2)$$

Имеются работы по поиску приближенных значений *F* [10, 12], которые в свою очередь могут быть использованы для получения приближенной версии $\mathrm{D}^{\mathrm{PML}}$ в (2), хотя, в общем случае, $\mathrm{D}^{\mathrm{MLE}} \neq \mathrm{D}^{\mathrm{PML}}$.

• Shlosser [13] выведен на основе предположения о перекосе::$\mathrm{E}[\mathrm{f_i}]/\mathrm{E}[\mathrm{f_1}] \approx \mathrm{F_i} / \mathrm{F_1}$, и хорошо работает, когда каждое отдельное значение появляется в среднем примерно один раз [14]. Он оценивает D как

$$D^{\text{Shlosser}} = d + \left(f_1 \sum_{i=1}^{n} (1-r)^i f_i\right) / \left(\sum_{i=1}^{n} ir(1-r)^{i-1} f_i\right) \qquad (3)$$

• Chao [15] аппроксимирует ожидаемую NDV, **E**[D], в большой популяции для некоторого базового распределения m, и оценивает NDV как нижнюю границу **E**[D] при приближении размера популяции к бесконечности:

$$D^{\text{Chao}} = d + f_1^2/(2f_2) \qquad (4)$$

• GEE [16] построен с использованием геометрического среднего для уравновешивания двух крайних случаев для значений, появляющихся точно один раз в выборке: значений с частотой один в C и попавших в S с вероятностью r v.s. значений с высокой частотой в C и хотя бы одним экземпляром, попавшим в S. Доказано, что он соответствует теоретической нижней границе ошибки отношения для оценки NDV в пределах постоянного фактора.

$$D^{\text{GEE}} = \sqrt{1/r} \cdot f_1 + \sum_{i=2}^{n} f_i \qquad (5)$$

• HYBGEE [16] - это гибридная оценка, использующая GEE для данных с большим перекосом и сглаженную оценку джек-ножа для данных с малым перекосом. AE [16] - это более принципиальная версия HYBGEE с плавным переходом от данных с малым перекосом к данным с большим перекосом.

2.5 Структура модели

Напомним, что нашей целью является обучение оценщика/модели для аппроксимации . Простые модели часто предполагают некоторую конкретную связь между входными характеристиками, т.е. профилем выборки в нашем случае, и меткой, которую нужно предсказать, т.е. NDV D.

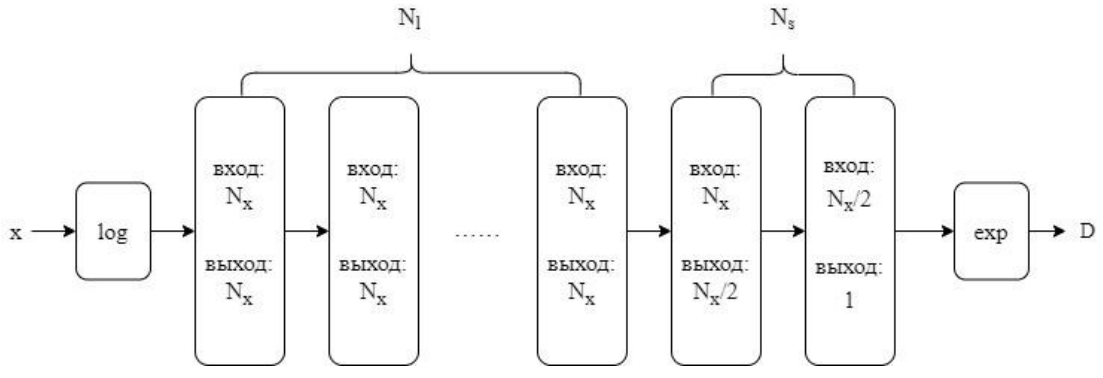

Рисунок 2 Архитектура сети: # Количество линейных слоев равно $N_l + N_s$.

Например, линейная/логистическая регрессия предполагает линейную зависимость между f и (преобразованным) D: например, $D = \sum_i a_i f_i$.Однако D^{MLE} может быть любой неизвестной функцией, которая намного сложнее, чем ограниченный выше класс линейных функций. Поэтому, поскольку нейронные сети способны аппроксимировать любую функцию (с достаточным количеством узлов и слоев) [17], мы выбираем нашу модель как нейронную сеть.

Заключение

В данном отчёте темой исследования является количество различных значений в графике машинного обучения.

В данной статье рассматриваются смежные работы по этой теме за последние годы, в том числе методы решения задач оценки NDV, и выбираются методы оценки NDV, основанные на выборке.

На основе сравнительного анализа предлагается основной вопрос данной работы: можно ли обучить модель методом машинного обучения, не зависящую от объема работы, для приближения к оценщику максимального правдоподобия (MLE)? Была проанализирована предварительная ситуация с формулой оценки ЧСС на основе MLE, а также рассмотрены некоторые репрезентативные оценки.

Проблема оценки ЧСС на основе выборок выражается как проблема MLE, однако, эта проблема даже трудно решаема приближенно. Была предложена схема "обучения-оценки" для обучения нейросетевой модели, которая не имеет отношения к рабочей нагрузке, чтобы приближенно оценить MLE, считая, что эта модель обучения выполнима.

Список использованной литературы

1. Ли Голян, Чжоу Сюаньхэ, Сунь Цзи, Юй Сян, Юань Хайтао, и Лю Цзяби н.Обзор технологий баз данных на основе машинного обучения. Журнал компьютерных наук, 43(11), 31,2020
2. Лейс, В., Радке, Б., Губичев, А., Кемпер, А., и Нойманн, Т. (2017, январь). Оценка кардинальности, выполненная правильно: выборка соединений на основе индексов. In Cidr.
3. Ахмед Метвалли, Дивьякант Агравал и Амр Эль Аббади. 2008. Зачем идт и логарифмическим путем, если можно идти линейным? К эффективному отчетливому подсчету поискового трафика. Труды 11-й международной к онференции по расширению технологий баз данных: достижения в област и технологий баз данных. 618-629.
4. Сяоин Ван, Чанбо Ку, Вэйюань Ву, Цзяньнань Ван и Цинцин Чжоу.2021. Готовы ли мы к обучаемой оценке кардинальности? Proc. VLDB Endow.14,9 (май 2021), 1640-1654.
5. Датт А., Ван С., Назн А. и др. Оценка селективности для предикатов диап азона с использованием облегченных моделей. Труды Фонда ВЛДБ, 2019, 12(9): 1044-1057.
6. Хилпрехт Б., Шмидт А., Кулесса М. и др. Deepdb: учитесь на данных, а не на запросах! Препринт arXiv arXiv: 1909.00607, 2019.
7. Чжу Р., Ву З., Хань Ю. и др. FLAT: быстрый, легкий и точный метод оцен ки количества элементов. Препринт arXiv arXiv: 2011.09022, 2020.
8. Рэймонд Л. Чемберс, Дэвид Г. Стил, Суоджин Ван и Алан Уэлш. 2012. Оц енка максимального правдоподобия для выборочных обследований. КПР Пресс.
9. Ву Р, Динг Б, Чу Х, и др. Учимся быть статистиком: обучаемый оценщик для числа отличных значений. препринт arXiv:2202.02800, 2022.
10. Чарикар М, Ширагур К, Сидфорд А. Эффективное профильное максимал ьное правдоподобие для оценки универсального симметричного свойства

// Труды 51-го ежегодного симпозиума ACM SIGACT по теории вычисле ний. 2019: 780-791.

11. Хао Й, Орлицкий А. Широкая оптимальность профильного максимальног о правдоподобия. Advances in Neural Information Processing Systems, 2019, 32.
12. Павличин Д.С., Цзяо Ж., Вайсман Т. Приближенное профильное максима льное правдоподобие. J. Mach. Learn. Res., 2019, 20: 122:1-122:55.
13. Шлоссер А. Об оценке размера словаря длинного текста на основе выбор ки. Инженерная кибернетика, 1981, 19(1): 97-102.
14. Haas P J, Naughton J F, Seshadri S, et al. Оценка числа различных значений атрибута на основе выборки//VLDB. 1995, 95: 311-322.
15. Чао А. Непараметрическая оценка числа классов в популяции. Скандинав ский журнал статистики, 1984: 265-270.
16. Чарикар М, Чаудхури С, Мотвани Р, и др. К гарантиям ошибок оценки дл я отдельных значений// Труды девятнадцатого симпозиума ACM SIGMOD-SIGACT-SIGART по принципам систем баз данных. 2000: 268-279.
17. Кубат М. Нейронные сети: всеобъемлющая основа Саймон Хайкин, Макм иллан, 1994, ISBN 0-02-352781-7. Обзор инженерии знаний, 1999, 13(4): 409-412.