УДК 004

Хаустова И.В.
студент группы ПИмд-2205а
Тольяттинский государственный университет
(г. Тольятти, Россия)

Научный руководитель:
Аникина О.В.
канд. тех. наук, доцент кафедры прикладной математики и информатики
Тольяттинский государственный университет
(г. Тольятти, Россия)

ВЛИЯНИЕ РАЗМЕРА ДАННЫХ НА ПРОИЗВОДИТЕЛЬНОСТЬ ПОЛНОСТЬЮ ГОМОМОРФНОГО ШИФРОВАНИЯ ДЛЯ ЗАЩИТЫ ДАННЫХ В ОБЛАЧНОМ МАШИННОМ ОБУЧЕНИИ

Аннотация: *в работе исследовано влияние размера данных на производительность полностью гомоморфного шифрования в облачном машинном обучении. На зашифрованных данных обучена модель логистической регрессии, решающая задачу прогнозирования задолженности по услугам ЖКХ. Данные для обучения модели разделены на выборки по разному количеству данных и признаков. Выбрана оптимальная конфигурация сервера для исключения влияния на процесс его производительности. Измерено и проанализировано время шифрования данных и обучения модели на наборах данных с разным количеством данных и количеством признаков.*

Ключевые слова: *полностью гомоморфное шифрование, TenSEAL, производительность полностью гомоморфного шифрования, логистическая регрессия, конфиденциальное машинное обучение.*

Введение

В настоящее время большинство современных компаний используют в своей деятельности машинное обучение, позволяющее решать различные

задачи: от прогнозирования задолженности до создания рекомендательных систем. При этом машинное обучение требовательно к вычислительной мощности, поэтому вместо закупки дорогостоящего оборудования часто используется арендованный облачный сервер или специального сервис, например, Yandex DataSphere. Но так как данные, использующиеся для обучения модели, передаются на сторонний сервер, возникает проблема обеспечения их безопасности, решить которую позволяет использование полностью гомоморфного шифрования, поддерживающего выполнение вычислений над зашифрованными данными, не раскрывая их содержание. Однако достаточно низкая производительность ограничивает применение полностью гомоморфных криптосистем, что связано с накоплением уровня шума, вычислительной сложностью алгоритмов шифрования и дешифрования, а также размером данных.

Целью данной работы является исследование влияния размера данных на производительность полностью гомоморфного шифрования в машинном обучении.

Для достижения поставленной цели в работе измеряется и анализируется время шифрования данных и время обучения модели логистической регрессии на зашифрованных данных, причём данные разделены на выборки по разному количеству данных и признаков.

Полностью гомоморфное шифрование данных с помощью библиотеки TenSEAL

В машинном обучении наиболее используемой библиотекой полностью гомоморфного шифрования является TenSEAL, например, данная библиотека использовалась в работах [2, 3, 4]. Данная библиотека с помощью API обеспечивает простоту использования языка Python, но реализует большинство операций на языке C++, что увеличивает быстродействие.

TenSEAL обеспечивает выполнение сложения, вычитания и умножения зашифрованных векторов. Для шифрования и дешифрования данных используется схема BFV в случае целых чисел и CKKS – в случае вещественных чисел.

Для шифрования данных в данной работе была использована схема CKKS, так как использование вещественных чисел позволяет достичь большей точности при обучении модели. Параметры шифрования данной схемы включают в себя степень полиномиального модуля и размеры модуля коэффициента.

При желаемом уровне защищенности, эквивалентном AES, равным 128 битам, и мультипликативной глубине модели логистической регрессии, равной 6, степень полиномиального модуля должна быть не ниже 8192, а модуль коэффициента должен содержать 8 простых чисел: первое и последнее по 40 бит и остальные по 21 бит [4].

Обучение модели логистической регрессии на зашифрованных данных

Для исследования факторов, влияющих на производительность полностью гомоморфного шифрования в облачном машинном обучении, в данной работе была разработана и обучена модель логистической регрессии, решающая задачу прогнозирования задолженности по услугам ЖКХ.

Логистическую регрессию можно рассматривать как простую однослойную нейронную сеть, использующую сигмоидальную функцию активации, поэтому модель была создана с помощью библиотеки PyTorch.

Так как в данной модели в качестве функции активации используется сигмоидная функция, то в связи с тем, что гомоморфные вычисления поддерживают только операции сложения и умножения, сигмоида была аппроксимирована полиномом третьей степени, полученным с помощью минимаксной аппроксимации, как показано работе [1].

Набор данных, используемый для обучения модели, содержит сведения о 4400 жителях, использующих цифровой сервис для расчетов в сфере ЖКХ, и включает 23 признака, такие как, например, тип жилья, сезонный фактор, подключенные уведомления или автоплатёж.

Выбор конфигурации облачного сервера

В качестве облачного сервера, используемого для обучения модели логистической регрессии, был выбран Yandex DataSphere – сервис для ML-разработки полного цикла, предоставляющий большой выбор готовых конфигураций вычислительных ресурсов.

Для исследования влияния размера данных на время обучения и шифрования необходимо исключить влияние фактора производительности сервера, для чего модель логистической регрессии была обучена на зашифрованных данных на доступных конфигурациях, как показано в таблице 1.

Таблица 1. Влияние конфигурации сервера на время шифрования и обучения модели логистической регрессии

Конфигурация сервера						**Время шифрования, с**	**Время обучения, с**
vCPU	**Количество vCPU**	**GPU**	**Количество GPU**	**RAM, ГБ**	**VRAM, ГБ**		
Intel Ice Lake	4	-	-	32	-	55	389
	8	-	-	64	-	54	388
Intel Broadwell	8	NVIDIA® Tesla® V100	1	96	32	54	388
	16		2	192	64	54	388

Как видно из таблицы 1, при достижении определенного уровня производительности сервера, его дальнейшее увеличение не ведёт к увеличению производительности полностью гомоморфного шифрования и не влияет на время шифрования и обучения, то есть достаточно использовать конфигурацию на платформе Intel Ice Lake с 8 ядрами и 64 Гб оперативной памяти.

Влияние объёма данных на время шифрования и обучения модели логистической регрессии

Для исследования влияния количества данных на время шифрования и обучения модели, из исходного набора данных было создано ещё 4 набора, в каждом

из которых было на 1000 объектов меньше, чем в предыдущем, то есть обучение проводилось на наборах с 4400, 3400, 2400, 1400 и 400 объектами соответственно. Результаты проведенного эксперимента приведены в таблице 2.

Таблица 2. Влияние количества данных на время шифрования и обучения модели логистической регрессии

Размер данных	Время шифрования, с	Время обучения, с
4400 x 23	54	388
3400 x 23	44	297
2400 x 23	29	219
1400 x 23	17	124
400 x 23	4	36

Как видно из таблицы 2, время шифрования и время обучения практически линейно зависят от количества данных, что также продемонстрировано на графиках на рисунках 1 и 2.

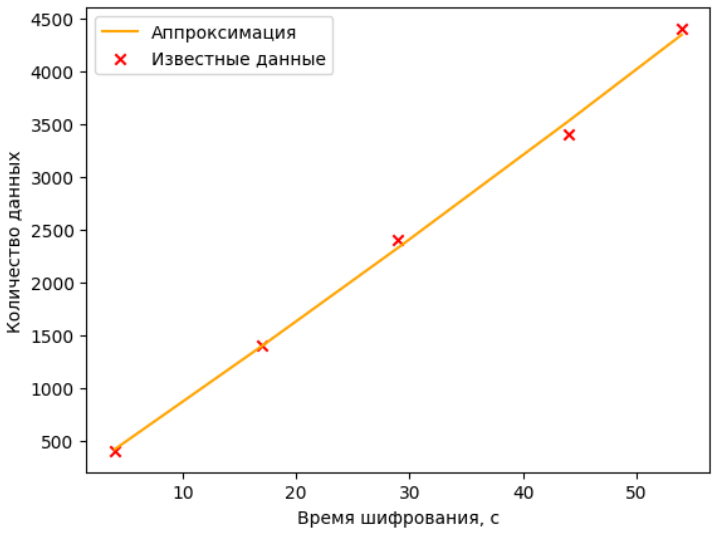

Рис. 1. График зависимости времени шифрования от количества данных

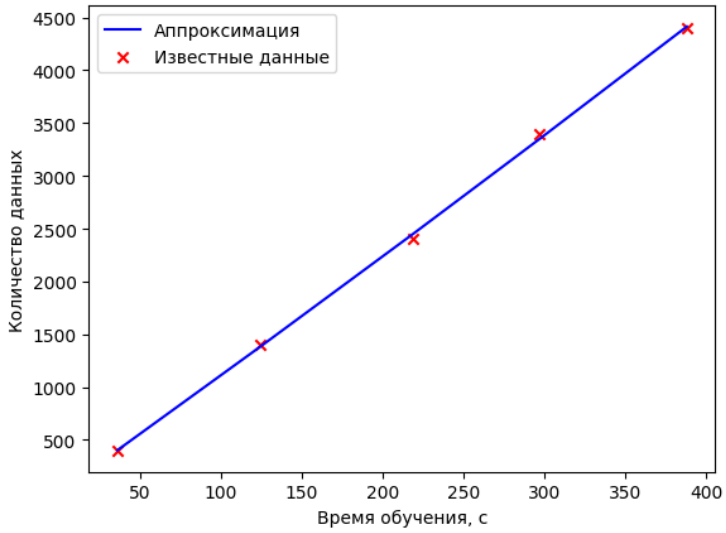

Рис. 2. График зависимости времени обучения от количества данных

Для исследования влияния количества признаков на время шифрования и обучения модели, из исходного набора данных было создано ещё 4 набора, в каждом из которых было на 2 признака меньше, чем в предыдущем, то есть обучение проводилось на наборах с 23, 21, 19, 17 и 15 признаками соответственно. Результаты проведенного эксперимента приведены в таблице 3.

Таблица 3. Влияние количества признаков на время шифрования и обучения модели логистической регрессии

Размер данных	Время шифрования, с	Время обучения, с
1400 x 23	17	124
1400 x 21	17	111
1400 x 19	17	104
1400 x 17	17	88
1400 x 15	17	75

Как видно из таблицы 3, время шифрования остается постоянным независимо от количества признаков, так как шифрование данных выполняется над всем набором данных целиком, а не отдельно для каждого признака. Таким образом, количество признаков не влияет на время шифрования, что также продемонстрировано на рисунке 3.

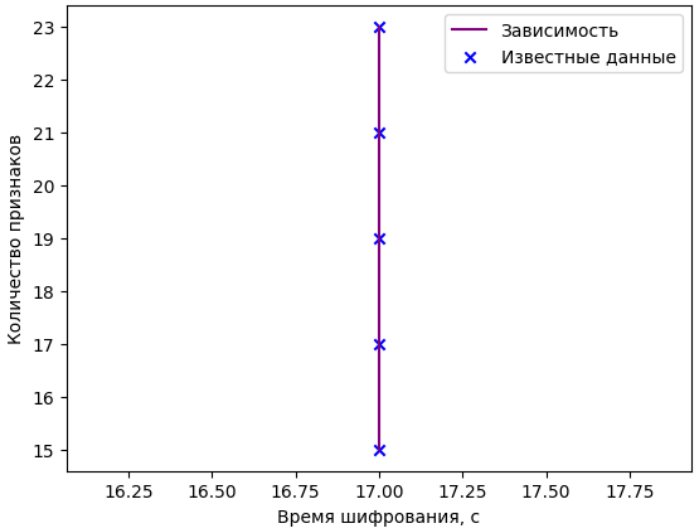

Рис. 3. График зависимости времени шифрования от количества признаков

Однако время обучения практически линейно уменьшается с уменьшением количества признаков, что связано с уменьшением сложности задачи оптимизации параметров модели, и что также продемонстрировано на рисунке 4.

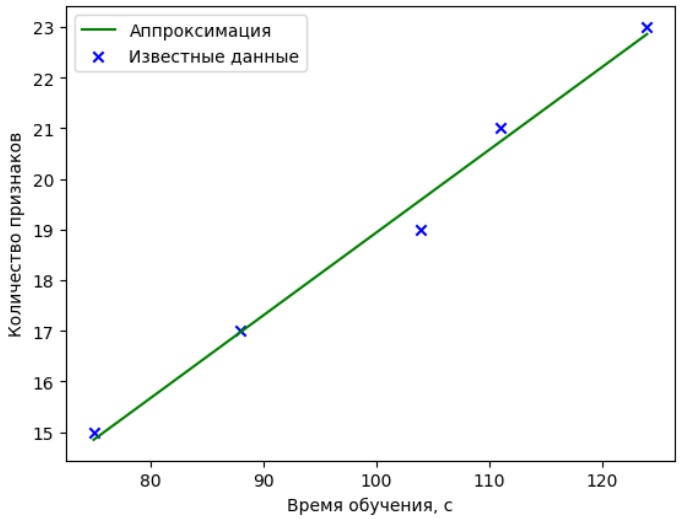

Рис. 4. График зависимости времени обучения от количества признаков

Таким образом, время шифрования данных растёт линейно при увеличении количества данных, но остаётся постоянным при увеличении количества признаков. В то время как время обучения растёт как при увеличении количества

данных, так и при увеличении количества признаков, при этом скорость изменяется значительно сильнее при переходе от большого количества данных и признаков к малому.

Заключение

В данной работе было исследовано влияние размера данных на производительность полностью гомоморфного шифрования, для чего модель логистической регрессии была обучена на зашифрованных данных, разделённых на выборки по разному количеству данных и признаков, после чего было измерено время шифрования данных и время обучения модели.

Для исключения влияния на эксперимент фактора производительности сервера была выбрана оптимальная конфигурация сервера на платформе Intel Ice Lake с 8 ядрами и 64 Гб оперативной памяти.

В результате эксперимента было установлено, что увеличение количества данных ведёт к пропорциональному росту времени шифрования и менее значительному росту времени обучения. В то же время уменьшение количества признаков не оказывает никакого влияния на время шифрования, но существенно снижает время обучения.

Таким образом, важен поиск баланса между безопасностью, необходимой точностью модели и скоростью обработки данных.

СПИСОК ЛИТЕРАТУРЫ:

1. Chen H, Gilad-Bachrach R., Han K., Huang Z., Jalali A., Laine K., Lauter K. Logistic regression over encrypted data from fully homomorphic encryption. – BMC Medical Genomics, Volume 11, Article number 81. – P. 56–67.

2. Mohassel P., Zhang Y. Secureml: A system for scalable privacy-preserving machine learning. – IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. – P. 19–38.

3. Vaikuntanathan C. J. V., Chandrakasan A. GAZELLE: a low latency framework for secure neural network inference [Электронный ресурс]. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/juvekar (дата обращения: 30.10.2024).

4. Хаустова И. В., Аникина О. В Использование полностью гомоморфного шифрования для защиты данных в машинном обучении в облаке // Вестник науки. 2024. № 4 (75). С. 482–491.

Khaustova I.V.
Togliatti State University
(Togliatti, Russia)

Anikina O.V.
Togliatti State University
(Togliatti, Russia)

IMPACT OF DATA SIZE ON THE PERFORMANCE OF FULLY HOMOMORPHIC ENCRYPTION FOR DATA PROTECTION IN CLOUD MACHINE LEARNING

***Abstract**: the paper investigates the impact of data size on the performance of fully homomorphic encryption in cloud machine learning. A logistic regression model is trained on encrypted data to predict utility service debt. The training data for the model is divided into samples with varying amounts of data and features. An optimal server configuration is selected to eliminate its impact on performance. The time for data encryption and model training is measured and analyzed on datasets with varying amounts of data and features.*

***Keywords**: fully homomorphic encryption, TenSEAL, performance of fully homomorphic encryption, logistic regression, confidential machine learning*