УДК 004.056

МОДИФИЦИРОВАННОЕ БЕЗОПАСНОЕ СКАЛЯРНОЕ ПРОИЗВЕДЕНИЕ В КОНФИДЕНЦИАЛЬНОМ КЛАСТЕРНОМ АНАЛИЗЕ K-MEANS С ВЕРТИКАЛЬНЫМ СЕКЦИОНИРОВАНИЕМ ДАННЫХ*

В. Г. Жуков, А. В. Вашкевич

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: zhukov.sibsau@gmail.com, alex23-5@yandex.ru

Рассматриваются актуальные проблемы в разработке и применении протоколов обеспечения конфиденциальности вычислений в многостороннем кластерном анализе данных. Приведено общее описание конфиденциальных многосторонних вычислений, определена модель нарушителя. Представлены недостатки одного из существующих решений конфиденциального кластерного анализа методом k-means при вертикальном секционировании данных; предложены пути их устранения с помощью корректного применения криптографического примитива, реализующего безопасное скалярное произведение. Приведённый математический аппарат описывает все аспекты исходного протокола и детали его модернизации. Статья содержит сравнительный анализ усовершенствованного криптографического примитива с аналогом – безопасной суммой. Сформулированы выводы об эффективности усовершенствованного криптографического примитива безопасного скалярного произведения и рекомендации по применению разработанной модели.

Ключевые слова: конфиденциальные многосторонние вычисления, криптографические примитивы, кластерный анализ k-means, вертикальное секционирование, безопасное скалярное произведение.

A MODIFIED SECURE DOT PRODUCT IN THE PRIVACY-PRESERVING K-MEANS CLUSTERING WITH THE VERTICAL PARTITIONING DATA

V. G. Zhukov, A. V. Vashkevich

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: zhukov.sibsau@gmail.com, alex23-5@yandex.ru

The authors consider the actual problems in the design and implementation of the privacy-preserving computations in the multiparty data clustering. There is a basic description of the secure multiparty computations and the definition of the adversary model in it. This article presents the flaws of existing solutions of the privacy-preserving k-means clustering over the vertically partitioned data, and proposed ways to eliminate these disadvantages by the correct using the cryptographic primitive realized secure dot product. The mathematical apparatus given in this article describes all aspects of the original protocol and the details of its modernization. The article has a comparative analysis of the advanced cryptographic primitive with its analog – secure sum. There are conclusions about the effectiveness of improved cryptographic primitive of the secure dot product and recommendations on application of the developed model.

Keywords: secure multiparty computations, cryptographic primitives, k-means clustering, vertical partition, secure dot product.

При проведении совместного анализа данных, принадлежащих разным организациям (пользователям) и являющихся конфиденциальными, не всегда достаточно обеспечения той конфиденциальности, что достигается при помощи традиционных криптосистем; необходимо обеспечить получение совместного результата без раскрытия самих исходных данных [1]. Задачи такого вида называются конфиденциальными многосторонними вычислениями (КМВ) – используются специальные протоколы, позволяющие нескольким участникам произвести вычисления на основе конфиденциальных входных данных каждого из них. После выполнения протокола КМВ каждый из участников получает результат вычисления, но ни один из них не должен иметь возможность получения никакой дополнительной информации о данных других участников.

В алгоритмах конфиденциальной кластеризации используется модель с получестными (semi-honest) участниками. Получестный участник следует протоколу, но волен использовать ту информацию, к которой он имеет доступ во время выполнения протокола, для попытки раскрытия данных других участников. При этом рассматривается возможность сговора (т. е. объединения знаний) некоторых из участников [2].

* Работа поддержана грантом Президента молодым кандидатам наук, договор № 14.124.13.473-МК от 04.02.2013.

Участник №1		Участник №2	Участник №3		
Атрибут №1	Атрибут №2	Атрибут №1	Атрибут №1	Атрибут №2	Атрибут №3

Рис. 1. Пример вертикального секционирования

На первом этапе исследований по применению КМВ в кластерном анализе рассматривался алгоритм, позволяющий работать с большими объемами данных и имеющий простую структуру. Наиболее распространенным и изученным алгоритмом кластерного анализа, удовлетворяющим предъявляемым требованиям, является алгоритм k-means.

Данные для многостороннего кластерного анализа могут быть по-разному распределены между участниками. Существует вертикальное (участники содержат разные столбцы данных), горизонтальное (участники содержат разные строки данных) и арбитральное (произвольное) секционирования данных. В вертикальном секционировании данных (рис. 1) каждый участник имеет набор только некоторых атрибутов, но каждого из объектов. При таком секционирование начальный выбор центров кластеров не является конфиденциальным, а защищаются такие шаги k-means, как вычисление расстояний между каждой точкой и центрами, выбор минимального расстояния для каждой точки, а также проверка условия остановки кластерного анализа.Алгоритм конфиденциальной кластеризации методом k-means при вертикальном секционировании:

Требуется: r участников, k кластеров, n точек.

1: для всех участников $j = 1, \ ..., \ r$

2: для всех кластеров $i = 1, \ ..., \ k$

3: произвольная инициализация центров кластеров μ'_{ij}

4: повторять

5: для всех участников $j = 1, \ ..., \ r$

6: для всех кластеров $i = 1, \ ..., \ k$

7: $\mu_{ij} = \mu'_{ij}$

8: кластер [i] = $\varnothing$

9: для всех точек $g = 1, \ ..., \ n$

10: для всех участников $j = 1, \ ..., \ r$

11: {Вычисление вектора X_j от точки g до каждого центра кластера по измерению j }

12: для всех кластеров $i = 1, \ ..., \ k$

13: x_{ij} = |данные точки g_j – данные центра кластера μ_{ij} |

14: каждый участник помещает g в кластер [ближайший] {алгоритм определения ближайшего кластера должен сохранять конфиденциальность}

15: для всех участников $j = 1, \ ..., \ r$

16: для всех кластеров $i = 1, \ ..., \ k$

17: μ'_{ij} – среднее всех точек кластера [i] по атрибуту j

18: пока не выполнится условие останова {алгоритм проверки условия останова должен сохранять конфиденциальность}

За основу для модернизации [3] взята схема конфиденциального кластерного анализа k-means при вертикальном секционировании, предложенная Саметом и соавторами [4], так как в отличие от других решений для данного секционирования в ней отсутствуют так называемые «особые» участники, выполняющие дополнительные операции с данными.

Обозначим набор атрибутов, содержащихся у P_i (участника под номером i), как $A_i = \{a_{i1}, \ a_{i2}, \ ..., \ a_{im}\}$. Также P_i владеет наборами атрибутов каждого из центров кластеров $\mu_{ji} = \{\mu_{ji1}, \ \mu_{ji2}, \ ..., \ \mu_{jim}\}$. Затем каждый участник суммирует расстояния между соответствующими измерениями. Так, порция расстояния между точкой и центром кластера μ_{ji} у участника P_i будет выглядеть следующим образом (пример для квадрата евклидового расстояния):

$$d_{ji} = \left(a_{i1} - \mu_{ji1}\right)^2 + \left(a_{i2} - \mu_{ji2}\right)^2 + ... + \left(a_{im} - \mu_{jim}\right)^2$$

Просуммировав по всем участникам эти расстояния, получим $d_{j1} + d_{j2} + ... + d_{jr}$, где r – число участников. Для центра другого кластера μ_q имеем похожую формулу $d_{q1} + d_{q2} + ... + d_{qr}$. Чтобы определить, какой из кластеров (j или q) ближе к точке, нужно просуммировать значения $\sum_{i=1}^{r}(d_{ji} - d_{qi}) = \sum_{i=1}^{r} d_i$. Если результат положительный, μ_q будет ближе к точке,

иначе ближе μ_j. Этот шаг повторяется $k-1$ раз, чтобы найти минимальное из расстояний от точки до центра кластера. При этом ни одно из расстояний не раскрывается.

Однако реализация данной схемы Саметом и соавторами имела два существенных недостатка, которые обязательно требовалось устранить:

– для двух участников применялся криптографический примитив «безопасное скалярное произведение» [5], предложенный Малеком и Мири;

– для участников количеством более двух применялся примитив «безопасная сумма» [6], в котором раскрываются данные любого из участников при сговоре его соседей по алгоритму.

Схема применения Саметом и соавторами криптографического примитива «Безопасное скалярное произведение»:

0) Участник P_1 имеет конфиденциальное значение $d_{j1} - d_{q1} = d_1$, а P_2 имеет конфиденциальное значение $d_{j2} - d_{q2} = d_2$, требуется получить числа l_1 и l_2 такие, чтоб l_1 хранился у P_1, l_2 у P_2 и $l_1 \cdot l_2 = d_1 + d_2$;

1) P_1 генерирует случайное ненулевое число l_1 и создаёт вектор $X = \left(\frac{d_1}{l_1}; \frac{1}{l_1}\right)$, а P_2 создаёт вектор $Y = (1; d_2)$;

2) P_1 и P_2 выполняют безопасное скалярное произведение [5] и P_2 получает число l_2 такое, что $l_1 \cdot l_2 = d_1 + d_2$.

3) P_2 посылает знак l_2. Если $l_2 = 0$, то до обоих кластеров одинаковое расстояние, если $l_2 \neq 0$, то знак показывает кластер, до которого расстояние меньше.

Формальное описание протокола безопасного скалярного произведения:

Предположим, что у двух участников (назовем их Алисой и Бобом) есть по *n*-мерному вектору: $\bar{x} = \{x_1, x_2, ..., x_n\}$ – у Алисы и $\bar{y} = \{y_1, y_2, ..., y_n\}$ – у Боба. Каждое измерение вектора является конечным полем F_p, а пространство векторов – конечным *n*-мерным расширением F_p^n конечного поля F_p.

Алиса и Боб хотят совместно и с сохранением конфиденциальности посчитать $\bar{x} \cdot \bar{y}$. Результат протокола должен быть известен только Алисе. Входные данные каждого из участников не должны быть раскрыты друг другу. Для достижения этих целей Алиса и Боб выполняют следующий протокол:

1. Алиса генерирует случайные вектор $\bar{\varphi} \in F_p^n$ и числа $a, b, c, d \in F_p$ такие, что $(ad - bc)^{-1} \neq 0 \in F_p$. Алиса вычисляет следующие векторы:

$$\bar{u} = a\bar{x} + b\bar{\varphi},$$

$$\bar{v} = c\bar{x} + d\bar{\varphi}.$$

2. Затем Алиса посылает векторы $\bar{u}$ и $\bar{v}$ Бобу. Зная свой вектор $\bar{y}$, Боб вычисляет $\bar{y} \cdot \bar{u}$ и $\bar{y} \cdot \bar{v}$. Затем Боб отправляет результаты обратно Алисе.

3. Алиса вычисляет $(ad - bc)^{-1}\left(d\left(\bar{y} \cdot \bar{u}\right) - b\left(\bar{y} \cdot \bar{v}\right)\right)$, что эквивалентно $\bar{x} \cdot \bar{y}$.

Следует отметить два момента, из-за которых применение данной схемы некорректно. Первым таким моментом является то, что примитив Малека и Мири предназначен для конечных полей. Таким образом, нельзя после выполнения примитива использовать знаки чисел l_1 и l_2 для достоверного определения знака суммы $d_1 + d_2$.

Пример, доказывающий данное утверждение:

Есть конечное поле

$$\{-5,\ -4,\ -3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3,\ 4,\ 5\}.$$

$$d_1 = -1,\ d_2 = -2.$$

Допустим, $l_1 = 2$. Тогда после проведения безопасного скалярного произведения будет получено $l_2 = 4$, так как в данном конечном поле $2 \cdot 4 = -1 - 2$. Знаки чисел l_1 и l_2 оба положительные, а знак суммы отрицателен. Следовательно, данный протокол не всегда работает в конечных полях.

Таким образом, использование знаков после операций в конечных полях не всегда будет давать корректный результат. Данный недостаток устраняется использованием множества вещественных чисел вместо конечных полей. При этом все операции безопасного скалярного произведения сохраняют корректность. Данное решение также имеет полезность в том, что использование конечных полей ухудшает точность проведения кластерного анализа, так как необходимо было бы «накладывать» область допустимых значений на конечное поле, что породило бы погрешность из-за необходимости квантования.

Вторым моментом, из-за которого некорректно применение предложенной Саметом и соавторами схемы, является использование двумерных векторов $(1; d_2)$ и $\left(\frac{d_1}{l_1}; \frac{1}{l_1}\right)$. Анализ криптографического примитива показал, что использование двумерных векторов раскрывает данные Боба Алисе. Причина в шаге 2 криптографического примитива, когда Боб посылает Алисе результаты скалярных произведений $\bar{y} \cdot \bar{u}$ и $\bar{y} \cdot \bar{v}$. Алиса, таким образом, получает два уравнения с двумя неизвестными:

$$y_1 u_1 + y_2 u_2 = \bar{y} \cdot \bar{u},$$

$$y_1 v_1 + y_2 v_2 = \bar{y} \cdot \bar{v},$$

где векторы $\bar{u}$ и $\bar{v}$ уже известны Алисе, а результаты скалярных произведений $\bar{y} \cdot \bar{u}$ и $\bar{y} \cdot \bar{v}$ она получила от Боба. Решив эти уравнения, Алиса вычисляет значе-

ния $\frac{d_1}{l_1}$ и $\frac{1}{l_1}$, откуда узнает d_1, т. е. косвенные сведения о данных Боба ($d_{j1} - d_{q1}$, разница между расстояниями от точки до одного кластера и до другого) становятся известны Алисе и, следовательно, нет обеспечения конфиденциальности данных одного из участников кластерного анализа.

Эту проблему можно решить, искусственно разбив два измерения векторов $(1; d_2)$ и $\left(\frac{d_1}{l_1}; \frac{1}{l_1}\right)$ на три. После разбиения векторы будут выглядеть следующим образом:

– вектор Алисы $(1; 1; d_2)$,

– вектор Боба $\left(\frac{d_1 + z}{l_1}; \frac{-z}{l_1}; \frac{1}{l_1}\right)$,

где z – случайное число, сгенерированное Бобом. Скалярное произведение данных векторов останется прежним, но теперь в двух уравнениях станет три неизвестных: $\frac{d_1 + z}{l_1}$, $\frac{-z}{l_1}$ и $\frac{1}{l_1}$, из которых «любопытному» (согласно модели нарушителя) участнику Алисе в общем случае будет нельзя получить значение d_1 Боба. Однако при определенных параметрах векторов $\bar{u}$ и $\bar{v}$ раскрытие информации все-таки останется возможным, поэтому Боб должен проверять присланные ему от Алисы векторы $\bar{u}$ и $\bar{v}$ на следующие условия (они могут выполняться по отдельности, но не одновременно):

$$u_1 = u_2,$$

$$v_1 = v_2,$$

потому что при выполнении обоих этих условий из уравнений можно будет узнать отдельно $\frac{d_1 + z}{l_1} + \left(\frac{-z}{l_1}\right)$ и $\frac{1}{l_1}$, а значит, получить $\left(\frac{d_1 + z}{l_1} + \left(\frac{-z}{l_1}\right)\right) \cdot l_1 = d_1$.

Также следует отметить, что безопасное скалярное произведение проводится для двух участников. При использовании этого примитива для нескольких участников следует применять схему, также описанную у Самета и соавторов. Пример для трёх участников:

1. P_3 дробит своё значение d_3 на случайные кусочки d_{31} и d_{32} такие, что $d_{31} + d_{32} = d_3$, и выбирает случайное число $r_3 \neq 0$.

2. P_3 и P_1 выполняют SDP, и P_1 получает s_1 такое, что $d_{31} + d_1 = r_3 \cdot s_1$.

3. P_3 и P_2 выполняют SDP, и P_2 получает s_2 такое, что $d_{32} + d_2 = r_3 \cdot s_2$.

4. P_2 и P_1 выполняют SDP для получения $s_1 + s_2$.

5. На выходе получается:

$$d_1 + d_2 + d_3 = r_3 \cdot (s_1 + s_2) = r_1 \cdot r_2 \cdot r_3$$ (рис. 2).

Причем участник P_i знает только d_i и r_i.

6. Если $r_1 = 0$, P_1 рассылает всем флаг, что расстояния до кластеров одинаковые, иначе он посылает знак r_1 участнику P_2, тот умножает знак r_1 на знак r_2 и посылает участнику P_3. Участник P_3 умножает полученный знак на знак r_3 и узнает, какой из кластеров ближе к точке, а затем рассылает всем.

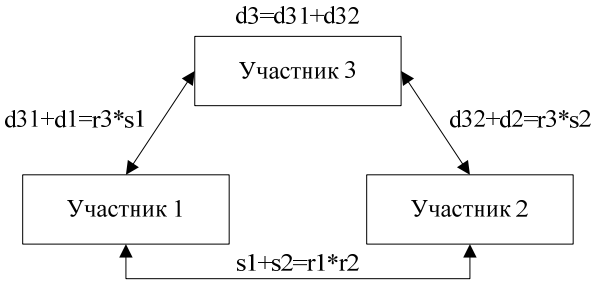

Рис. 2. Схема безопасного скалярного произведения для трех участников

Основной минус данного алгоритма в том, что количество данных, передаваемых по сети, растёт пропорционально квадрату количества участников. Именно по этой причине алгоритм не подходит для кластерного анализа с большим количеством участников и обрабатываемых данных, поскольку он применяется на каждой итерации для каждого объекта данных $k - 1$ раз.

В связи с проектированием программного обеспечения, реализующего данный протокол, стало возможно сравнить сетевой трафик (рис. 3), генерируемый алгоритмом с использованием того или иного криптографического примитива.

Тем не менее, если к конфиденциальности предъявляются жёсткие требования, следует применять примитив «безопасное скалярное произведение», поскольку в нём гарантируется сохранение конфиденциальности даже при сговоре против одного из участников всех остальных (рис. 4).

Таким образом, устранив недостатки схемы по применению безопасного скалярного произведения, можно использовать данный криптографический примитив для определения ближайшего кластера. Однако применение безопасного скалярного произведения требует передачи данных от каждого участника каждому, что при большом количестве участников и большом количестве объектов данных приведет к большому объему трафика. Поэтому использование безопасного скалярного произведения рекомендуется к применению для малого количества объектов данных или участников, когда велик риск сговора и на первый план выходит обеспечение конфиденциальности данных в ущерб скорости работы алгоритма.

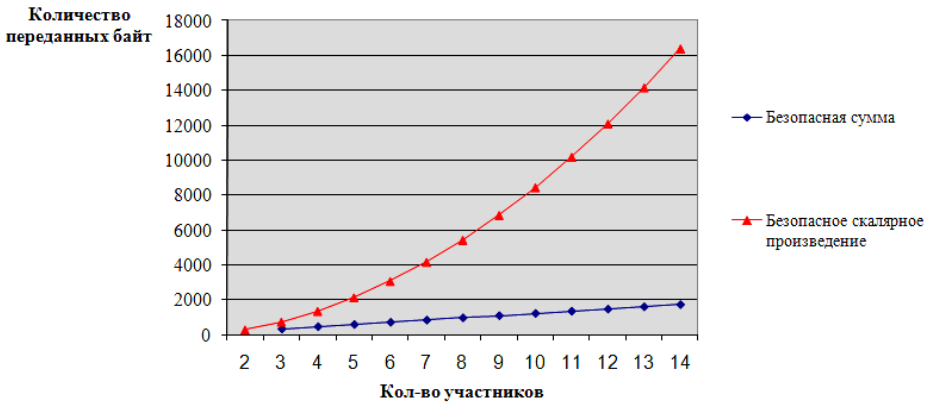

Рис. 3. Зависимость количества переданных байт за одну операцию сравнения расстояний от количества участников

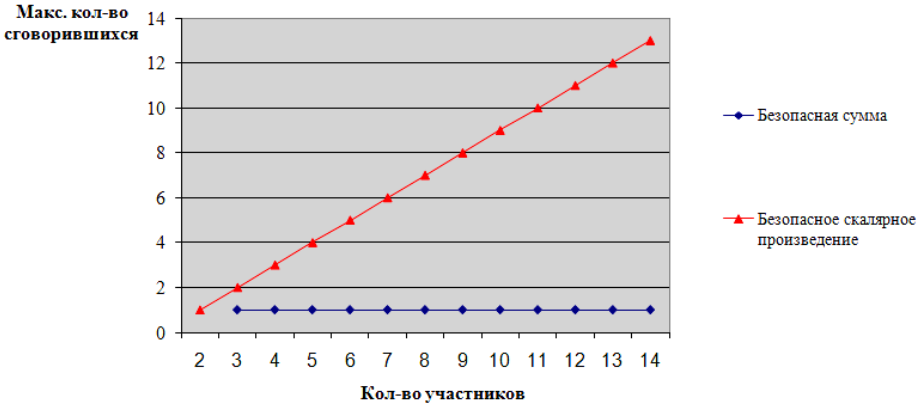

Рис. 4. Зависимость максимального допустимого количества сговорившихся участников от общего количества участников

Библиографические ссылки

1. Шутый Р. С. Рандомизированные протоколы, применяемые для выполнения конфиденциальных многосторонних вычислений в компьютерных сетях / Санкт-Петербургский гос. ун-т телекоммуникаций им. проф. М. А. Бонч-Бруевича, 2009. 170 с.

2. Жуков В. Г., Стефаров А. П. Формирование типовой модели нарушителя правил разграничения доступа в автоматизированных системах // Известия ЮФУ. Техн. науки: науч.-техн. и прикладной журнал. Таганрог : ТТИ ЮФУ, 2012. Вып. 12 (137). С. 45–54.

3. Жуков В. Г., Вашкевич А. В. Конфиденциальный кластерный анализ при вертикальном секционировании данных // Информационная безопасность – 2013 : материалы XIII Междунар. науч.-практ. конф. Ч. I. Таганрог : ТТИ ЮФУ, 2013. С. 191–198.

4. Samet S., Miri A., Orozco-Barbosa L. Privacy-Preserving K-Means Clustering in Multi-Party Environment // Proceedings of International Conference on Security and Cryptography, Barcelona, Spain, 2007. С. 381–385.

5. Malek B., Miri A. Secure dot-product protocol using trace functions // Proceedings of the 2006 IEEE International Symposium on Information Theory, 2006. С. 927–931.

6. Clifton C., Kantarcioglu M., Vaidya J., Lin X., Zhu M. Tools for privacy preserving data mining // SIGKDD Explorations, 2002. Vol. 4(2). С. 28–34.

References

1. Shutyi R. *Randomizirovannye protokoly, primenjaemye dlja vypolnenija konfidencial'nyh mnogostoronnih vychislenij v komp'juternyh setjah* (Randomized protocols applied for performing secure multiparty computations in computer networks). The Bonch-Bruevich Saint-Petersburg State University of Telecommunications, 2009, 170 p.

2. Zhukov V. G., Stefarov A. P. *Izvestija Juzhnogo federal'nogo universiteta. Tehnicheskie nauki*, 2012, no. 12 (137), p. 45–54.

3. Zhukov V. G., Vashkevich A. V. *Materialy XIII mezhdunarodnoj nauchno-prakticheskoj konferencii "Informacionnaja bezopasnost'-2013"* (Proceedings of the 13th International Scientific Conference "Information Security-2013"), Russia, Taganrog, 2013, vol. 1, p. 191–198.

4. Samet S., Miri A., Orozco-Barbosa L. *Proceedings of the 2007 International Conference on Security and Cryptography, Barcelona, Spain*, 2007, p. 381–385.

5. Malek B., Miri A. *Proceedings of the 2006 IEEE International Symposium on Information Theory*, 2006, p. 927–931.

6. Clifton C., Kantarcioglu M., Vaidya J., Lin X., Zhu M. Tools for privacy preserving data mining (2002) *SIGKDD Explorations*, 4(2), p. 28–34.