https://doi.org/10.31891/2219-9365-2022-69-1-2
УДК 004

Володимир ФЕРЕНС
Хмельницький національний університет

ВИКОРИСТАННЯ АЛГОРИТМУ ЗБУРЕННЯ ДАНИХ ДЛЯ ЗАХИСТУ КОНФІДЕНЦІЙНОСТІ

В роботі проведено дослідження алгоритмів захисту конфіденційності для великомасштабних даних на основі машинного навчання. У сучасному світі у повсякденному житті та для дослідників і практиків проблема захисту конфіденційності великих даних є чи не найосновнішою проблемою та актуальною задачею. За результатами проведених досліджень удосконалено метод захисту конфіденційності для більш ефективного використання енергоносіїв. Також набув подальшого розвитку метод, за допомогою якого забезпечується можливість імплементації його у систему охорони здоров'я.

Ключові слова: конфіденційність, безпека даних, алгоритм захисту конфіденційності, аналітика даних і машинне навчання, Інтернет речей,конденсація даних, адаптивний шум .

Volodymyr FERENS
Khmelnitsky National University

THE USE OF A DATA-EXCITATION ALGORITHM TO PROTECT CONFIDENTIALITY

Combining a large number of different technologies, such as the Internet of Things, cloud computing, computational computing and machine learning, contributes to the rapid and active spread of technological development in various fields, such as health, energy, agriculture, transportation, etc. The increase in the number of publicly available gadgets has contributed to the rapid growth of the Internet of Things, becoming one of the main sources of large data flows.

Cyberspace covers not only the physical sphere, but also the human, a huge amount of information (data) becomes available for analysis. Analytical processing of large amounts of data, with the development of their generation speed, gives excellent results and provides extreme accuracy in creating important ideas. One of the approaches that attracts the most attention is in-depth training, which provides high performance with large amounts of data. Various areas, including the above, work closely with data privacy. They show a tendency to increase the consequences due to the disclosure of confidential data to third parties, attacks on databases and more. Corporations and other organizations are constantly striving to ensure maximum confidentiality and that all information is stored on the company's local servers. To prevent breaches of privacy, machine learning, together with data analytics, should implement all possible privacy protection scenarios to ensure that users' privacy is not compromised. There are many approaches to maintaining confidentiality, however, the sheer size of large amounts of data and data flows make maintaining a confidentiality a challenge. The main problem among the existing approaches is their inability to maintain the right balance between confidentiality, usefulness and efficiency when dealing with large amounts of data. Some effective approaches provide good privacy but do not provide sufficient performance during data operations, while others, on the contrary, provide good performance but do not provide a high level of privacy.

This paper investigates privacy protection algorithms for large-scale data based on machine learning. In today's world in everyday life and for researchers and practitioners, the problem of protecting the privacy of big data is the most basic problem and urgent challenge. Based on the results of this research, a privacy protection method has been improved for more efficient use of energy. The method has also been further developed to enable its implementation in the healthcare system.

Key words: Confidentiality, Data security, Data analytics and machine learning, Internet of Things(IoT), Data Condensation, adaptive noise.

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ТА ПРАКТИЧНИМИ ЗАВДАННЯМИ

Об'єднання великої кількості різних технологій, таких як Інтернет речей, хмарні обчислення, граничні обчислення та машинне навчання сприяють стрімкому та активному поширенню технологічного розвитку у різних сферах, таких як охорона здоров'я, енергетика, агропромисловість, транспортування тощо. Збільшення кількості загальнодоступних гаджетів сприяло швидкому зростанню Інтернету речей, ставши одним з основних джерел великих потоків даних.

Кіберпростір охоплює не тільки фізичну сферу, але і людську, величезна кількість інформації (даних) стають доступними для аналізу. Аналітична обробка великих об'ємів даних, з розвитком швидкості генерації їх, дає чудові результати та забезпечують надзвичайну точність у створені важливих ідей. Одним із підходів, що привертає найбільшу кількість уваги є поглиблене навчання, яким забезпечується висока продуктивність з обсягами великих даних. З конфіденційністю даних тісно співпрацюють різні сфери, зокрема вищеназвані. У них помітна тенденція зростання наслідків через відкриття конфіденційних даних третім особам, атак на бази даних тощо. Корпорації та інші організації постійно прагнуть, забезпечення максимальної конфіденційності та щоб вся інформація зберігалась на локальних серверах компанії. Для запобігання порушень конфіденційності машинне навчання разом із аналітикою даних повинно запровадити всі можливі сценарії захисту конфіденційності для забезпечення неушкодженості конфіденційності

користувачів. Існує багато підходів до збереження конфіденційності, однак, величезні розміри великої кількості даних і потоків даних роблять збереження конфіденційності важкою задачею. Головною проблемою серед існуючих підходів є їх неспроможність підтримки правильного балансу між конфіденційністю, корисністю та ефективністю при оперуванні з великою кількістю даних. Деякі ефективні підходи забезпечують хорошу конфіденційність, але не забезпечують достатньої швидкодії під час операцій із даними, у той час як інші – навпаки забезпечують хорошу швидкодію, але не забезпечують високий рівень конфіденційності.

Постійні вдосконалення у сфері аналітики даних і технік машинного навчання, таких як поглиблене навчання, довели, що дають високу точність, забезпечуючи надійні прогнози. Однак такі підходи часто покладаються на величезні обсяги даних, які, можливо, потрібно буде збирати із різних джерел. Для прикладу, моделі які тренуються на масиві бази даних, зазвичай, розкривають особисту інформацію, та є вразливими до атак, які ставлять за мету порушити конфіденційність даних. Окрім цього, вразливою, є область аналізу біометричних даних: це можуть бути як і данні із датчика аналізу відбитків пальців, що є чи не в кожному смартфоні, так і розпізнавання обличь та сканування сітківки ока. Ці області виконують ресурсозатратні та важкі завдання, які часто залучають сторонні сервера, а отже до яких можуть отримати доступ зловмисники. Тому, якщо біометрична інформація неконтрольовано доставляється на ненадійні та сумнівні сторонні сервери, то це можна вважати чималим витоком конфіденційності (тобто неконтрольований витік інформації), оскільки данні отримані із біометричних датчиків можна співвідносити із конфіденційними даними записів медперсоналу та інформації банківських операцій. Атаки на конфіденційність намагаються зробити все, для виявлення ідентичності осіб у вхідних даних, які використовуються для кожного екземпляру даних або моделі машинного навчання. Таким чином, порушення конфіденційності може значно вразити цілісність та безпеку інформації, а також передати особисту інформацію в ненадійне середовище, до якого кібер-зловмисники звертаються зі зловмисним наміром.

АНАЛІЗ ЛІТЕРАТУРНИХ ДАНИХ ТА ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Кількісно визначити конфіденційність конкретних методів конфіденційності за допомогою показників конфіденційності досить складно. У літературі наведено [1] велику кількість визначень показників конфіденційності. Відповідно до конкретних типів методів конфіденційності (наприклад, втручання в збурення багатовимірних даних). Таким чином можна показати показник конфіденційності, який є методом збурення даних основою якого є адаптивний шум, що базується на залежності від близькості між порушеним значенням і його вихідним значенням. Цей показник оцінює вірогідність [c] вихідної оцінки протягом інтервалу часу [a, b]. Наразі конфіденційність оцінюється за інтервалом [a, b] та вірогідністю [c]. Цей метод має проблему неприйняття. Запропоновано розподіл бази вихідних даних на рахунки. На основі ентропії даних існує більш загальний метод кількісної оцінки конфіденційності. Поняття «інформація, що пов’язана з даними» використовується для визначення рівня конфіденційності. Однак одним із основних недоліків цього підходу є те, що він не враховує фактори, які викликають ризики, що є напряму пов’язані із змагальними атаками. Для додаткового методу інтерференції даних на основі шуму з урахуванням близькості між початковими значеннями та вихідні значеннями. Одним із основних цього підходу є те, що він так чи інакше не враховує ризики, які є напряму пов’язані з атаками зловмисника на основі базових знань. У цьому сценарії для обміну даними конфіденційності існують різні атрибути. І передбачено, що в них присутні різні рівні захисту конфіденційності. Вищесказане несе більш конфіденційну інформацію, ніж інші атрибути і також тісно пов’язано з ними. Між іншим, різниця присутня як і між збуреними та і між невиправленими атрибутами. Основна важкість – це оцінка початкових даних після використання збурення. Вища різниця дисперсії забезпечує[2-4] відповідну її конфіденційність. А статистична різниця (названої дисперсії) передбачає більш високий рівень складності при аналізі вхідних даних, а саме їх оцінці. Покращення рівню конфіденційності (найслабшого атрибута) є основною метою збереження конфіденційності. Однак підходи до збереження конфіденційності, які використовують цей підхід часто не є ефективними. Така продуктивність зумовлена ітераційним характером пошуку оптимальної цілі для покращення конфіденційності атрибута, що володіє найменшою силою. Для подолання невизначеного рівня приватності була введена модель конфіденційності[4-5]. Моделі конфіденційності включають: анонімність, багатоманітність, стислість і диференційну конфіденційність. Модель конфіденційності гарантує концептуальний підхід до виконання та дотримання суворих умов безпеки. Отже модель конфіденційності під час глибокого аналізу даних і машинного навчання забезпечує достатньо організований спосіб кількісної оцінки конфіденційності.

Є три технологічні підходи конфіденційності: контроль над розкриттям статистичних даних, інтелектуальний аналіз даних для збереження конфіденційності і технології, що підвищують конфіденційність. Шифрування на основі атрибутів, контроль доступу за допомогою аутентифікації, контроль доступу на основі часу та розташування, а також використання протоколів на основі обмежень – це деякі механізми, які використовуються для покращення конфіденційності систем у динамічних середовищах. Основні підходи для інтелектуального аналізу даних для збереження конфіденційності можна

розмежувати за чотирма типами: криптографічні підходи, збурення даних, незбурювані підходи, штучне генерування даних. Серед різноманіття підходів до інтелектуального аналізу даних часто надають перевагу збуренню (модифікації) даних через його простоту, ефективність та можливість налаштування компромісу між конфіденційністю. Ці властивості роблять збурення даних найкращою альтернативою, як, для прикладу, великі дані та потоки даних.

Коли справжні дані (вихідні) повністю змінюються перед тим, як вони повідомляються третій стороні, яка потенційно може бути зловмисником – збурення даних. Порівнюючи з гомоморфним шифруванням, або із аналогічним криптографічним підходом, змінені дані не піддаються жодним форматам перетворення, зважаючи на те, що просторова та часова важкість є меншою. Оскільки гомоморфні шифрування[4] забезпечують достатньо високий рівень конфіденційності. Але слід враховувати, що під час збурення даних рівень витоку даних хоч і часто мінімальний, але присутній, тому слід передбачувати всі можливі виходи даних, щоб не було порушень цілісності даних чи їх конфіденційності. Початкове значення x згенеровується шляхом додавання до x випадкової величини r до x або шляхом застосування певного процесу рандомізації до x. На рисунку 1 показано збурення даних, котрі можна розділити на два класи: 1) збурення вхідних даних та 2) збурення вихідних даних.

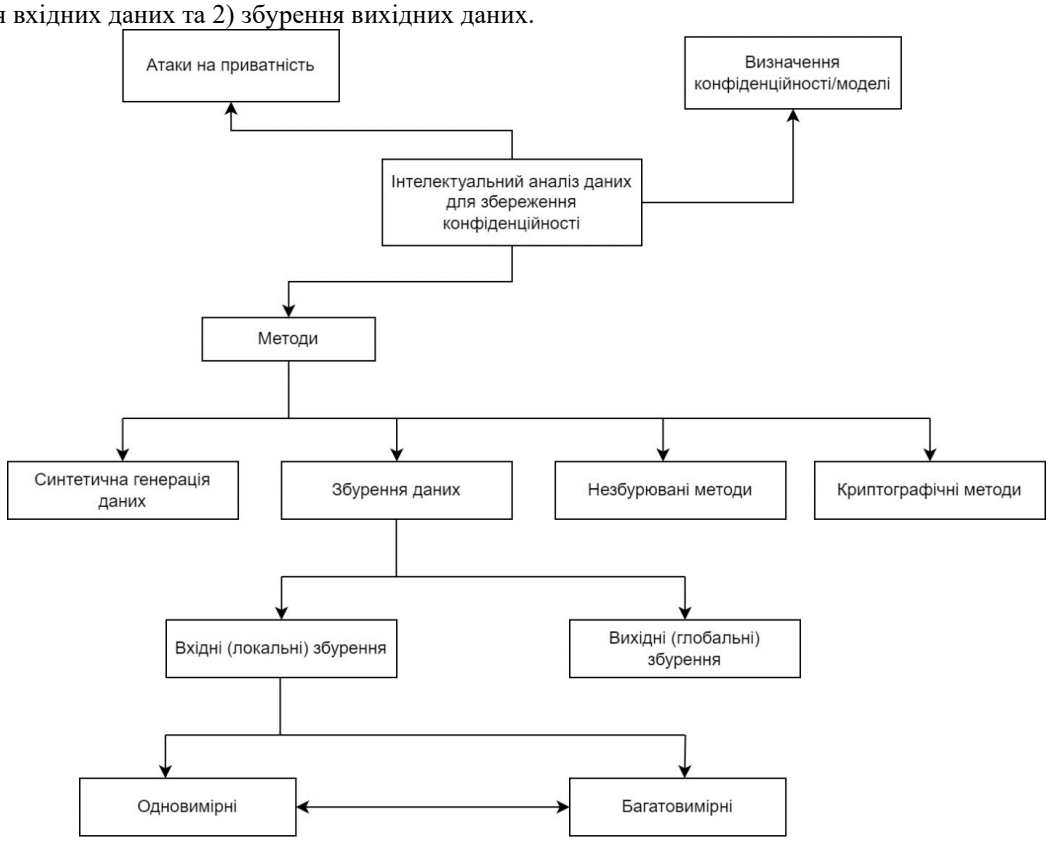

Рис. 1. Класифікація існуючих методів інтелектуального аналізу даних для збереження конфіденційності

Виходячи із цих методів запропоновано систему збереження конфіденційності, як підхід(алгоритм) до розподіленого машинного навчання. Цей алгоритм є розподіленим алгоритмом, що сприяє збереженню конфіденційності, який використовує збурення даних. Алгоритм підтримує збурення великих даних із підтримкою їх поширення між розподіленими об’єктами без порушення конфіденційності. Фактичне збурення відбувається в розподілених сутностях локальних пристроїв включаючи локальні правила та глобальні параметри збурень. Тому, алгоритм збурення розподілених даних обмежує вихідні дані для передачі (до збурення) через мережу, через необхідність захисту інформації від третіх сторін, які можуть несанкціоновано отримувати доступ до даних, та діяти із ними у своїх, злочинних цілях для отримання власної вигоди. Генерація глобальних параметрів збурення інформації із високою імовірністю може гарантувати відсутність погіршення точності чи стійкості до атаки збурених даних. Як наслідок алгоритм успішно запобігає витоку конфіденційності.

Важливим є використання метрики конфіденційності із багатьма колонками для оцінки запропонованого методу. Алгоритм забезпечує розподіл на користувачів, власників чи адміністраторів наборів даних відповідно до їх ролей у тій чи іншій сфері (рисунок 2), де ці дані застосовуються відповідно правил. Це забезпечують і додаткові релізи версій вихідного набору даних. Отже, вихідний набір буде всього-на-всього недоступним для користувачів, і не зможе бути доступним за жодних обставин та причин.

Рис. 2. Приклад розподіленої організаційної структури

На рисунку 2 зображено екосистема охорони здоров'я[5] яка територіально розподілена між кількома місцями. Система охорони здоров'я може мати безліч розподілених філій, що сприяють та збирають безліч даних про охорону здоров'я, включаючи дані датчиків інтернету речей. Центральний орган у переважній більшості випадків координує роботу розподілених лікарень з погляду з погляду підтримки цілісності даних підтримки широкого спектра аналітики. Центральний орган (дослідний центр) також відповідає за обмін даними з третіми сторонами для підвищення інтелектуальності та якості обслуговування пацієнтів.

Збурення вихідних даних називається глобальним збуренням даних, тоді як збурення вхідних даних -- локальним збуренням даних. На рисунку 1 показано, що збурення даних проводиться на даних, коли вони полишають власників даних. Довірена особа при збурені даних (переважно вихідних) застосовує збурення даних до вихідних даних, які отримуються в результаті виконання запитів на них. Не дивно, що регулярність застосування збурень як вхідних так і вихідних даних, що потребують конфіденційності є досить високою. Таким чином збурення переважно варто застосовувати на збурені вхідних даних, саме для ненадійних середовищ, де треті особи мають доступ до інформації, або безпека середовища не є на високому рівні. Якщо порівнювати вхідне і вихідне збурення даних, то на вхідних – присутнє застосування, якого неможливо уникнути вищого рівня випадковості, що дає забезпечення і посилення рівнів конфіденційності, ніж збурення власне вихідних даних.

Для розширення вказаної ідеї розглянуто дисперсію різниці між збуреними і незбуреними наборами даних. Атрибут, який розглянуто у цьому методі повертає мінімальну дисперсію для різниці, та розглядається як мінімальна гарантія збереження конфіденційності даних. Якщо X^pвідображає збурений ряд даних атрибуту X, то рівень конфіденційності методу збурення варто визначати:

$$Var(P), \text{ при } P = (X^p - X). \tag{1}$$

В такому випадку має місце рівність:

$$Var(P) = Var(p_1, p_2, \dots, p_n) = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})^2. \tag{2}$$

Найкращі параметри збурення визначаються наступним чином. Під час кожної ітерації відбувається максимізація значення φ. Це власне і сприяє створеню значення ϕ. Як зазначено у рівняннях нижче. При цьому вісь відбиття змінюється від 1 до n (кількості атрибутів). Кут повороту при цьому збмінюється у діапазоні значень від 0° до 179° Таким чином повертається найбільше значення ϕ для мінімальної гарантії конфіденційності φ.

$$\phi = \max\left([[\varphi_j]_{j=1}^{179}]\right). \tag{3}$$

Наступним кроком є генерація матриці обертання за допомогою найкращих параметрів збурення. АЗК записує оптимальний кут повороту і вісь відбиття на Ф. Наступне рівняння слід використовувати відповідно оптимального кута.

$$RF_{\overline{ND}} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}_{(n+1)(n+1)} \quad (4)$$

Далі слід застосувати складене перетворення відображення, переміщення та обертання до нормалізованої матриці з використанням матриць оптимального відображення та оптимального обертання.

Обґрунтування та технічна новизна. Алгоритм захисту конфіденційності для великих даних застосовує геометричні перетворення із достатньо оптимальними та оптимізованими параметрами збурення[6] при цьому збільшує випадковість за допомогою випадкових розширень та відповідних перестановок кортежів. Він визначає конфіденційність так, що результуючий набір даних має високий рівень різниці, який є оптимальним у порівнянні з вихідним набором даних. Це все відповідно нашого розподілу[7], що використовується в алгоритмі, оскільки відбувається мінімізація простору пошуку, та знаходження найкращої моделі можливого збурення даних, можна стверджувати, що вона є корисною для конкретного набора даних.

Оцінку продуктивності зосереджено на корисності, яку можна визначити як зручність використання або як ефективність використання збурення даних. Досліджено корисність алгоритму з точки зору класифікації. Таким чином алгоритм захисту конфіденційності ставить за мету довести комфортність використання, яка заснована на тимчасовій складності, боротьбу із надмірними витратами пам'яті, масштабованості та упередженнях, що охоплюють основу оцінки. Вихідні набори даних були збурені за допомогою алгоритму, геометричним та обертальним збуренням[8-9]. Тому, це дало можливість порівняти результати за допомогою непараметричного статистичного тесту. Обчислювана складність алгоритму полягає у виконуваній центральній сутності, що відповідає O(k) для кількості екземплярів, де k – постійна змінна.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Сьогодні існує велике різноманіття систем, які є сучасними та прогресивними. Для прикладу може слугувати банківська справа чи вищезгадана охорона здоров'я. Ці системи є часто обмежені. Зокрема це відбувається у аналітичному використанні належних механізмів для обміну даними із відповідним забезпеченням конфіденційності для аналітики. Оскільки наш алгоритм є композиційним та його без вагань може бути розширено для нарощення його функціоналу. Він може забезпечити конфіденційність машинного навчання, яке є розподіленим. У запропонованій системі весь контроль генерації даних та глобальних параметрів збурення належить центральному контролюючому органу, в той час як локальне збурення даних може бути проведене генерацією глобальних параметрів.

ЛІТЕРАТУРА

1. From old to new: Assessing cybersecurity risks for an evolving smart grid / L. Langer, F. Skopik, P. Smith, M. Kammerstetter. *Computers & Security*. 2016. № 62. P. 165–176

2. Chen, H.-L., Doty, D., Soloveichik, D.: Deterministic function computation with chemical reaction networks. *Nat. Comput*. 13, 517–534 (2014)

3. Censor-Hillel, K., Parter, M. & Schwartzman, G. Derandomizing local distributed algorithms under bandwidth restrictions. *Distrib. Comput.* 33, 349–366 (2020). URL:https://doi.org/10.1007/s00446-020-00376-1

4. Методики визначення вихідних даних для оцінки залишкових ризиків при забезпеченні конфіденційності інформаційних об'єктів URL: http://drsp.ipri.kiev.ua/article/view/142921 Privacy preserving data classification with rotation perturbation, *Fifth IEEE International Conference on Data Mining (ICDM'05)* 27-30 Nov. 2005 URL: https://ieeexplore.ieee.org/abstract/document/1565733

5. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys* Volume 51 Issue 4 July 2019 Article No.: 79 pp 1–35 URL:https://dl.acm.org/doi/abs/10.1145/3214303

6. The Enhancement of Security in Healthcare Information Systems *Journal of Medical Systems* volume 36, p.1673–1688 (2012) URL: https://link.springer.com/article/10.1007/s10916-010-9628-3

7. Homotopy perturbation method with an auxiliary parameter for nonlinear oscillators Journal of Soft *Computing Paradigm (JSCP)* (2021) Vol.03/ No.01 Pages: 19-28 URL: https://journals.sagepub.com/doi/full/10.1177/1461348418811028

8. Big Data Analysis and Perturbation using Data Mining Algorithm *Journal of Soft Computing Paradigm (JSCP)* (2021) Vol.03/ No.01 p. 19-28 URL: https://irojournals.com/jscp/V3/I1/03.pdf

9. Generalized random rotation perturbation for vertically partitioned data sets, IEEE Symposium on *Computational Intelligence and Data Mining*. 2019 IEEE Xplore: 15 May 2009 URL: https://ieeexplore.ieee.org/abstract/document/4938644

REFERENCES

1. From old to new: Assessing cybersecurity risks for an evolving smart grid / L. Langer, F. Skopik, P. Smith, M. Kammerstetter. Computers & Security. 2016. № 62. P. 165–176

2. Chen, H.-L., Doty, D., Soloveichik, D.: Deterministic function computation with chemical reaction networks. Nat. Comput. 13, 517–534 (2014)

3. Censor-Hillel, K., Parter, M. & Schwartzman, G. Derandomizing local distributed algorithms under bandwidth restrictions. Distrib. Comput. 33, 349–366 (2020). URL: https://doi.org/10.1007/s00446-020-00376-1

4. Methods for determining the source data for the assessment of residual risks in ensuring the confidentiality of information objects URL: http://drsp.ipri.kiev.ua/article/view/142921 Privacy preserving data classification with rotation perturbation, Fifth IEEE International Conference on Data Mining (ICDM'05) 27-30 Nov. 2005 URL: https://ieeexplore.ieee.org/abstract/document/1565733

5. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. *ACM Computing Surveys* Volume 51 Issue 4 July 2019 Article No.: 79 pp 1–35 URL: https: //dl.acm.org/doi/abs/10.1145/3214303

6. The Enhancement of Security in Healthcare Information Systems Journal of Medical Systems volume 36, p.1673–1688 (2012) URL: https://link.springer.com/article/10.1007/s10916-010-9628-3

7. Homotopy perturbation method with an auxiliary parameter for nonlinear oscillators Journal of Soft Computing Paradigm (JSCP) (2021) Vol.03 / No.01 Pages: 19-28 URL: https://journals.sagepub.com/doi/full /10.1177/1461348418811028

8. Big Data Analysis and Perturbation using Data Mining Algorithm Journal of Soft Computing Paradigm (JSCP) (2021) Vol.03 / No.01 p. 19-28 URL: https://irojournals.com/jscp/V3/I1/03.pdf

9. Generalized random rotation perturbation for vertically partitioned data sets, IEEE Symposium on Computational Intelligence and Data Mining. 2019 IEEE Xplore: 15 May 2009 URL: https://ieeexplore.ieee.org/abstract/document/4938644