Biomedical Data Science
2025, VOL. 5, NO. 1, 42-44
DOI: 10.47297/wspbdsWSP2752-630509.20250501

Application and Effect Evaluation of Differential Privacy in Genomic Data Medical Services

Yan Yu

Nursing College, Jiujiang University, Jiujiang, Jiangxi, 332005, China

ABSTRACT
Genomic data is increasingly widely used in the medical service field, bringing new opportunities for medical development. However, the issue of privacy protection has become prominent, hindering further applications. Differential privacy technology provides a new idea to solve this problem, as it can maintain data availability while ensuring privacy. This article analyzes the threats of genomic data privacy leakage, covering sequencing storage, research analysis, and clinical medical service processes. It elaborates on the differential privacy method, including its formal definition, indistinguishability and other characteristics, and evaluates the application effect in combination with research cases such as pharmacogenomics and personalized medicine. Currently, in practical applications, there are still challenges such as balancing privacy and data utility, technology scalability, and privacy budget setting. In the future, this technology will develop in the direction of integration and innovation with other technologies, expanding application scenarios, and realizing standardization and normalization, and is expected to provide stronger privacy protection support for precision medicine.

KEYWORDS
Differential privacy; Genomic data privacy; Privacy protection; Medical services; Application evaluation

1 Introduction

With the rapid development of biotechnology, genomic data is increasingly widely used in medical services, bringing revolutionary changes to fields such as precision medicine, disease prevention, and personalized treatment. However, due to the high sensitivity of personal information, the privacy protection of genomic data has become more and more prominent. How to effectively utilize genomic data to promote the development of medical services while ensuring data security has become an important issue that needs to be solved urgently. Computer cryptography experts Bonnie Berger and Sean Simmons proposed that differential privacy technology can be used to protect genomic databases and prevent the leakage of personal genomic data[1]. As an emerging privacy protection technology, differential privacy adds controlled random noise to the data, providing a new solution for the privacy protection of genomic data while protecting data privacy and maintaining the statistical characteristics of the data as much as possible.

2 Threats of Genomic Data Privacy Leakage

2.1 Genomic Data Sequencing and Storage

Under the current legal supervision, sequencing centers are trustworthy. However, some human factors cannot be ignored. For example, sensitive data may be leaked due to commercial interests. Similarly, data centers are also trustworthy, but unpredictable risk factors cannot be excluded from interfering with the data. Data centers obtain raw data from sequencing centers for storage and management. During data transmission and storage, privacy data may be leaked due to malicious attacks, equipment failures and other problems. Therefore, it is necessary to desensitize during the sequencing process and protect during transmission and storage.

2.2 Genomic Data Research and Analysis

Genome-wide association studies (GWAS) are one of the most common types in the research on the association between genotypes and diseases. In data collection, GWAS usually requires a large number of samples to determine reproducible genome-wide significant associations. One of the questions researchers need to consider is how to protect the security of these data. At the same time, as more and more research institutions participate in human genomics research, the privacy protection requirements of individuals involved in data collection are also receiving more and more attention.

2.3 Genomic Data in Clinical Medical Services

With the continuous development of genomic data research, genomic data has been applied to many aspects of

clinical medical services such as genetic diagnosis, genetic disease screening, drug development, and paternity testing. Due to the fact that genomic data is related to sensitive issues such as genetics, health, and family, there are privacy protection problems in medical services. To address this issue, differential privacy and cryptography methods are currently used to protect the data privacy of genomic data in clinical medical services.

3 Differential Privacy Method and Application Effect

3.1 Differential Privacy Method

The differential privacy model was first proposed by Cynthia Dwork in 2006 [2]. Its purpose is to ensure that the impact of individual data on statistical results is negligible, so as to prevent attackers from inferring individual information through statistical results. Its core idea is that for two datasets that differ by one record, the probabilities of obtaining the same result when querying them are very close. The measure of "very close" is determined by the privacy budget ε. The smaller the ε, the stronger the privacy. For example, when an attacker queries 1000 pieces of information in the information database and the other 999 pieces of information after removing any one piece of information, if the obtained query results are the same, the attacker cannot determine the removed piece of information. In this case, we say that the individual corresponding to this piece of information has obtained privacy protection.

The formal definition of differential privacy is as follows: Given a random algorithm M, for any two neighboring datasets D and D′ (that is, D and D′ differ by only one record), if the algorithm M satisfies:

$$\Pr\left[M(D) \in O\right] \leqslant e^{\epsilon} \Pr\left[M(D') \in O\right] + \delta$$

Then the algorithm M is said to satisfy (ε,δ)-differential privacy protection. When $\delta = 0$, the algorithm is said to satisfy ε-differential privacy.

After being perturbed by the random algorithm M, the probabilities of obtaining a specific output O should differ little. The parameter ε is the privacy protection budget, which is used to control the privacy protection strength. The smaller the ε value, the stronger the privacy protection, but the accuracy of the data will also be affected.

3.2 Main Characteristics of the Differential Privacy Method

Indistinguishability: Differential privacy can ensure the indistinguishability of individual privacy information, that is, an attacker cannot distinguish whether the sensitive information of a certain individual is included in the published data.

Controllable privacy protection strength: By adjusting the size of the noise (that is, the privacy budget ε), the strength of privacy protection can be controlled.

Randomization mechanism: Differential privacy ensures data privacy through the introduction of a randomization mechanism. The introduction of noise makes it impossible for an attacker to accurately infer personal information.

3.3 Application and Evaluation of Differential Privacy in Genomic Medical Services

Fredrikson et al. used a differential privacy machine learning model to protect patient privacy in pharmacogenomics research and guide medical treatment based on patients' genotype and background information [3]. Specifically, Laplace noise was added to the coefficients of the linear regression model to achieve the purpose of privacy protection. In the study of personalized Warfarin dosage, this method can effectively achieve privacy protection. However, it should be noted that although differential privacy can be used for simulated clinical trial analysis, its application may have a significant impact on utility.

In personalized medicine, the existing differential privacy learning methods cannot improve the prediction of feasible data size and dimension. Honkela et al. used a differential privacy linear regression model to achieve privacy protection in drug sensitivity prediction[4]. In this model, noise is added to the statistics derived from the data to ensure privacy. In this process, the Wishart mechanism is used to perturb the covariance terms of the input data, and the Laplace mechanism is used to perturb the target terms, so as to achieve the purpose of differential privacy. This method is an asymptotically consistent and efficient privacy protection method and has wide applicability and can be extended to other predictors. In the case of limited data, the sharing of restricted privacy information is reduced, and projection technology is used to deal with outliers to adapt to more stringent boundary conditions. This strategy reduces the amount of noise required to achieve the same level of privacy protection, thus effectively improving the data utility under limited data.

In recent years, scholars have proposed a series of differential privacy-based feature selection techniques, such as differential privacy random forest (Random Forest) and reusable reservation set (Reusable Reservation Set). However, in the field of bioinformatics, the number of features far exceeds the number of observation samples, and the application of differential privacy technology often faces the risk of overfitting. To address this challenge, Le et al. designed and used a

differential privacy-based Evaporative Cooling method to achieve privacy protection in individual disease classification [5]. They randomly select attributes using the exponential mechanism, combine the Relief - F algorithm for feature screening, and use the random forest model to perform privacy-protected classification tasks, thus effectively avoiding the overfitting problem.

4 Current Challenges and Future Prospects

4.1 Challenges Faced

Although differential privacy has shown significant application effects in genomic data medical services, it still faces some challenges: Balancing privacy and data utility: The introduction of differential privacy technology will reduce the utility of data, affecting subsequent data analysis, research, and decision-making. How to maximize the retention of data utility while protecting individual privacy is one of the current research difficulties. Scalability of differential privacy technology: Genomic data is large and diverse, requiring differential privacy technology to be able to handle the processing and protection of large-scale data. Therefore, it is necessary to further improve the scalability of differential privacy technology to adapt to the characteristics and requirements of genomic data.

Setting of privacy budget: The setting of the privacy budget ε has an important impact on the effect of differential privacy technology. How to reasonably set the value of the privacy budget ε according to specific application scenarios and data characteristics to achieve the best balance between privacy protection and data accuracy is an important research direction at present.

4.2 Future Prospects

With the continuous development of gene sequencing technology and the accumulation of medical big data, the application of genomic data in medical services will be more extensive. Differential privacy, as an effective privacy protection technology, will play an increasingly important role in the privacy protection of genomic data. In the future, the application of differential privacy technology in genomic data medical services will show the following trends:

Technology integration and innovation: Differential privacy technology will be integrated with other privacy protection technologies, data mining technologies, and machine learning technologies to form more efficient and accurate privacy protection schemes. Expansion of application scenarios: Differential privacy technology will be applied to more genomic data medical service scenarios, such as disease prediction, personalized medicine, and drug development, providing stronger support for precision medicine. Standardization and normalization: With the wide application of differential privacy technology in genomic data medical services, relevant standards and norms will gradually be formed to guide the practical application and evaluation of differential privacy technology.

5 Conclusion

This article discussed the application and effect evaluation of differential privacy in genomic data medical services. The application analysis proved the effectiveness of differential privacy in protecting the privacy of genomic data. Differential privacy technology ensures that individual privacy is not leaked by adding random noise to the data while maintaining the statistical characteristics of the data. However, differential privacy still faces some challenges in genomic data medical services, such as balancing privacy and data utility, scalability, and setting of privacy budget.

References

[1] Lin C., Wang P., Song H., Zhou Y., Liu Q., & Wu G. (2016). A differential privacy protection scheme for sensitive big data in body sensor networks. Annals of Telecommunications, 71, 465 - 475.

[2] Dwork C. (2006). Differential Privacy. International Colloquium on Automata, Languages and Programming.

[3] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proc USENIX Secur Symp. 2014 Aug;2014:17-32. PMID: 27077138; PMCID:PMC4827719.

[4] Liu H., Peng C., Wu Z., Tian Y., & Tian F. (2021). Review of theories and methods for genomic data privacy protection. Chinese Journal of Computers (07), 1430 – 1480.

[5] Le T.T., Simmons K.W., Misaki M., Bodurka J., White B.C., Savitz J.B., & McKinney B.A. (2017). Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. Bioinformatics, 33, 2906–2913.