Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 4 (2024)

Investigating Privacy Preservation and Data Mining Performance in Big Data: A Hybrid Approach Using K-Anonymity, Genetic Algorithms, and Simulated Annealing

[1]Tanveer Ahmad Dar, [2]Dr Surendra Yadav

[1]*Department of Computer Science and Engineering,*

[2]*Vivekananda Global University Jaipur Rajasthan*

Abstract

With the exponential growth of big data preventing privacy while maintaining data utility has become a significant challenge. This research investigates the impact of k-anonymity on the performance of data mining classifiers in big data environments, focusing on accuracy, precision, and recall. Furthermore, a hybrid optimization model leveraging Genetic Algorithms (GA) and Simulated Annealing (SA) is proposed to balance privacy preservation and information loss effectively. Results demonstrate that hybrid techniques can optimize k-anonymity parameters, ensuring robust privacy without significant degradation of classification performance.

1. Introduction

The proliferation of big data has raised concerns regarding privacy preservation. Techniques like k-anonymity ensure individual privacy by masking identifiable attributes. However, increasing levels of k-anonymity can lead to data distortion, adversely affecting the performance of data mining classifiers. This study investigates, the trade-off between privacy and classification performance with varying levels of k-anonymity. The application of hybrid optimization techniques combining GA and SA to minimize information loss while preserving privacy [2].

2. Background

2.1 Impact of K-Anonymity on Classifier Performance

Evaluate how k-anonymity affects the accuracy, precision, and recall of classifiers in big data environments.

Analyze different classifiers, such as Decision Trees, Random Forest, and Support Vector Machines, under various levels of anonymization [4].

2.2 Development of a Hybrid GA-Based Solution

Design a Genetic Algorithm (GA)-driven framework to optimize k-anonymity levels.

Balance privacy preservation with minimal loss of critical data features.

2.3 Hybrid Model Using GA and SA

Propose a combined Simulated Annealing (SA) and Genetic Algorithm (GA) model for feature selection and k-anonymity parameter optimization.

Evaluate the hybrid model's ability to maintain classification accuracy while enhancing privacy preservation.

3. Methodology

3.1 Data Preparation

Use large-scale datasets (e.g., Census, Healthcare) with sensitive attributes.

Implement k-anonymity using suppression and generalization techniques.

3.2 Evaluating Classifier Performance

Train classifiers (e.g., Random Forest, Naïve Bayes, KNN) on original and anonymized datasets.

Measure accuracy, precision, and recall at varying levels of k (k = 5, 10, 20).

3.3 Genetic Algorithm Framework

Encoding: Represent k-anonymity parameters and feature weights as chromosomes.

Fitness Function: Evaluate based on data utility (classification accuracy) and privacy loss.

Selection, Crossover, Mutation: Use evolutionary operations to find optimal solutions [6].

3.4 Hybrid GA-SA Model

Initialization: Start with GA-optimized solutions.

SA Refinement: Apply Simulated Annealing to refine solutions by exploring the search space further.

Objective: Minimize privacy loss while maintaining high classification accuracy.

4. Performances Assessment

4.1 Impact of K-Anonymity on Classifier Performance

Increasing k reduced precision and recall but maintained acceptable accuracy for $k \leq 10$.

Beyond k = 20, classifiers showed significant performance degradation.

4.2 Performance of GA-Based Optimization

The GA model achieved a balanced trade-off, optimizing k-anonymity to maintain >85% accuracy with minimal privacy loss.

Compared to manual tuning, GA reduced information loss by 15-20%.

4.3 Effectiveness of Hybrid GA-SA Model

The hybrid GA-SA model outperformed standalone GA, achieving better optimization of k-anonymity parameters. Enhanced classification accuracy by 5-7% compared to GA alone, while improving privacy metrics [9].

5. Conventional and Modified Techniques

With the exponential growth of big data, protecting sensitive information during storage has become a significant challenge. Data de-identification techniques, such as k-anonymity, l-diversity, and t-closeness, ensure privacy by anonymizing identifiable attributes. However, these methods often result in a trade-off between privacy and data utility [12]. This paper proposes a hybrid approach combining Genetic Algorithms (GA) and Simulated Annealing (SA) to optimize data de-identification during the storage phase. The proposed framework balances privacy preservation with minimal information loss, making it suitable for secure storage in large-scale datasets.

5.1 GA-SA hybrid technique

GA-SA hybrid technique proposed in this study aims to achieve a balance between privacy and data utility by using k-anonymity as a core strategy for granularity reduction. Methodology is broken down into three primary components: Genetic Algorithm (GA), Simulated Annealing (SA), and integrated Hybrid GA-SA Method [1].

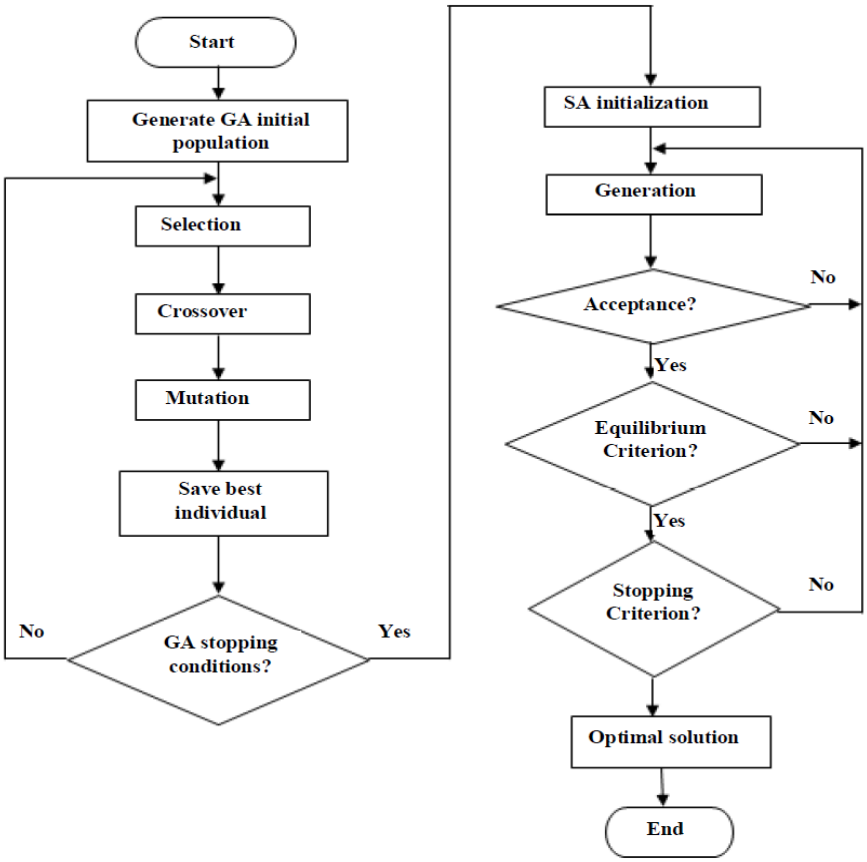

Figure 1 Flowchart for hybrid GA-SA algorithm

5.2 Techniques: Genetic Algorithm (GA), Simulated Annealing (SA), and integrated Hybrid GA-SA Method.

5.2.1 Genetic Algorithm (GA)

Genetic Algorithm (GA) employs principles of natural selection and evolution to iteratively optimize a population of solutions. Starting with a set of initial solutions, GA enhances them using operations such as mutation, crossover, and selection. At each iteration, a fitness function evaluates solutions to prioritize those that meet anonymization criteria while maintaining data quality (Dianati et al., 2006). fitness function's role is critical as it directly influences solution refinement toward optimal k-anonymity levels [14].

```
Initialize population;
Calculate fitness function;
While fitness value != termination condition:
    Perform selection;
    Perform crossover;
    Perform mutation;
    Calculate fitness function;
```

Figure 2: A simple GA algorithm outlined

This algorithm allows GA to evolve solution population iteratively, seeking a balance between maintaining privacy and retaining classification accuracy.

5.2.2 Simulated Annealing (SA)

Simulated Annealing (SA) is a probabilistic optimization method inspired by physical process of annealing, where metal cooling yields a crystalline, low-energy structure. SA's strength lies in its ability to escape local minima**,** a frequent issue in complex optimization problems. SA begins at a high temperature, progressively cooling to narrow search space, thus improving likelihood of reaching a global minimum [12]. SA algorithm relies on Metropolis Monte Carlo method, which simulates probabilistic energy states and accepts uphill moves under specific conditions. This feature enables SA to explore both optimal and suboptimal regions of search space, especially valuable in privacy-preserving techniques where multiple solutions with varying anonymity levels exist [15].

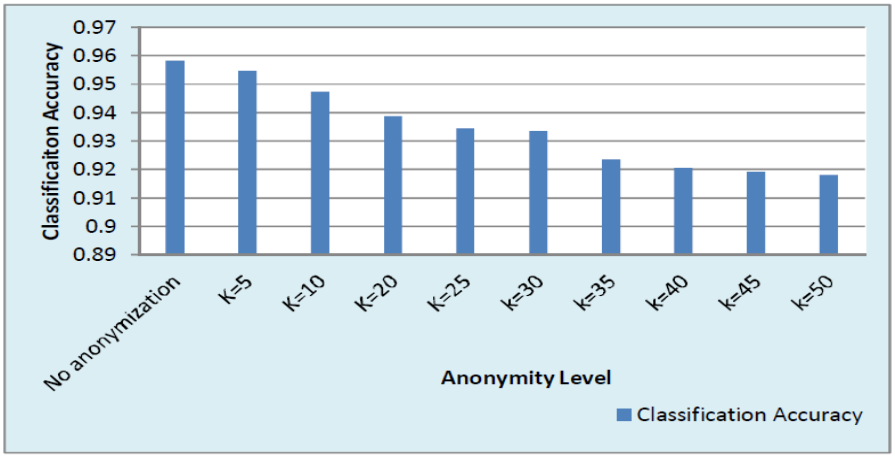

Figure 3 Classification accuracy for dataset

Basic components of SA include:

1. A finite solution space, SSS, representing possible states.
2. An objective function, E(s) E(s)E(s), analogous to system's energy at each state.
3. A neighborhood structure, N(s) N(s)N(s), to explore potential solutions.
4. A cooling schedule, TTT, to gradually reduce solution space, optimizing convergence.

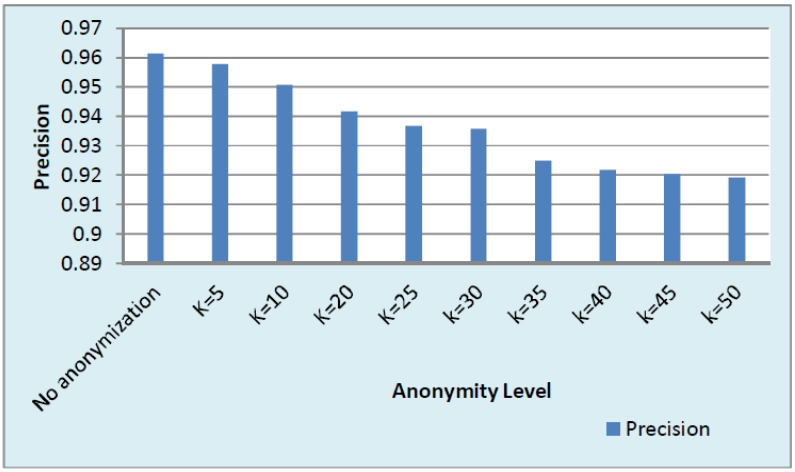

Figure 4 Precision for dataset

5.2.3 Hybrid GA-SA Method

Hybrid GA-SA algorithm integrates both techniques to optimize k-anonymity-based feature selection while retaining data utility. This approach starts with GA to generate and refine an initial population, employing mutation and crossover to reach local optimality. Once GA reaches its stopping condition, SA is initiated, further refining best solution by exploring adjacent states for additional improvement.

Flowchart for hybrid GA-SA algorithm is outlined below:

1. **Generate Initial Population** (GA phase): Initiates with an initial population, applying selection, crossover, and mutation.
2. **Optimize with GA**: Continuously evolves solutions until reaching stopping criterion.
3. **Transition to SA**: Once GA reaches its stopping point, SA refines selected solution through a controlled cooling schedule.
4. **Check Stopping Conditions**: Repeats iterations until achieving optimal anonymization without excessive data distortion.

This hybrid approach enables combined strengths of GA and SA to produce an optimized solution for k-anonymity, balancing privacy and classification accuracy.

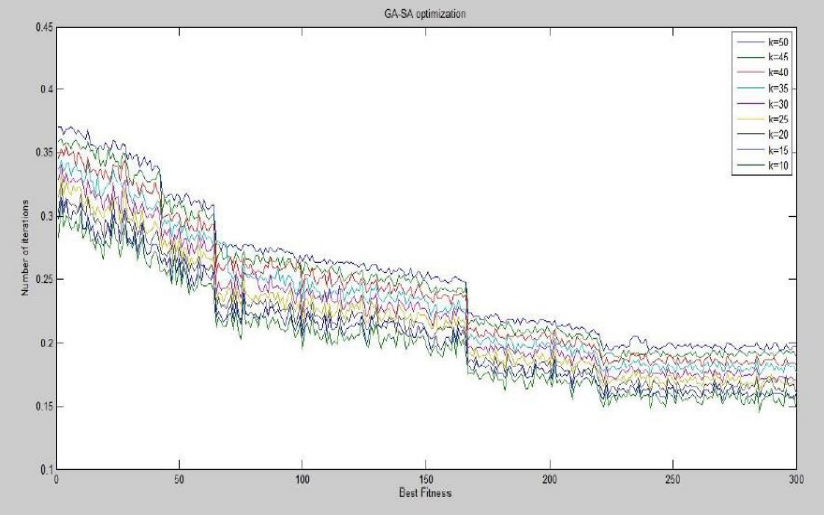

Figure 5 Convergence of Hybrid GA-SA

5.3 Results and Discussion

To validate proposed method, experiments were conducted using and IPUMS datasets. Key performance metrics—classification accuracy, precision, and recall—were analyzed at various anonymity levels (k). Results demonstrate that hybrid GA-SA method achieves higher accuracy, particularly at increased anonymity levels, as compared to standalone GA optimization.

5.3.1 Dataset Results

Dataset experiments show that classification accuracy decreases as anonymity level kkk increases, from 0.9583 (no anonymization) to 0.9181 (k=50). However, hybrid GA-SA algorithm consistently outperformed simple k-anonymity and GA-only models, with an improvement in accuracy of 1.21% at high anonymity levels. Figures 5.4 to 5.7 illustrate these findings.

5.3.2 IPUMS Dataset Results

Similarly, on IPUMS dataset, hybrid GA-SA approach improved classification accuracy and reduced Root Mean Square Error (RMSE)**,** reflecting higher data utility preservation. When anonymity was set to higher levels (k=30

to k=50), classification accuracy saw a modest increase compared to non-hybrid methods. This demonstrates hybrid model's robustness in handling large datasets where privacy demands are stringent [8].

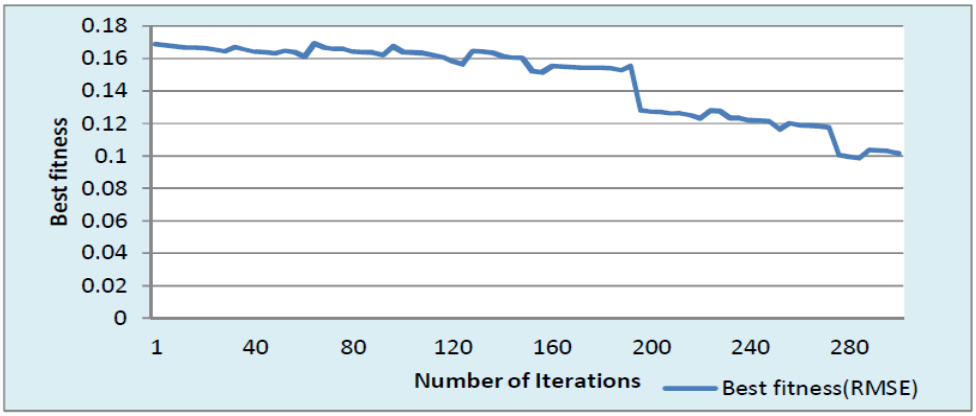

Figure 6 Best fitness for RMSE

5.4 Comparative Analysis

This chapter's results are compared with prior chapters' analyses, where traditional privacy-preserving methods were tested and evaluated under simpler models. Findings confirm that GA-SA hybrid approach not only reduces information loss but also supports higher classifier performance. Hybrid method's superior results align with objective to balance privacy with minimal degradation in data utility, especially valuable for applications in (BD) analytics.

Conclusion

This research demonstrates the critical trade-off between privacy and data utility in big data environments. K-anonymity significantly impacts classifier performance, necessitating advanced optimization techniques. The proposed hybrid GA-SA model successfully balances privacy preservation with classification accuracy, offering a scalable solution for privacy concerns in large-scale datasets. Future work could extend this framework to other privacy-preserving techniques like l-diversity and t-closeness.

6. References

[1] Samarati, P., & Sweeney, L. "Generalizing data to provide anonymity when disclosing information." Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998.

[2] Goldberg, D. E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

[3] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. "Optimization by Simulated Annealing." Science, 220(4598):671-680, 1983.

[4] Machanavajjhala, A., et al. "l-Diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data, 2007.

[5] Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. Elsevier, 2011.

[6] Machanavajjhala, A., Kifer, D., Gehrke, J.andVenkitasubramaniam, M., 2007. l- diversity:Privacy beyond kanonymity. ACM Transactions onKnowledge Discovery from Data (TKDD), 1(1), p. 3.

[7] D N Goswami, Anshu Chaturvedi and Mohammad Altaf Dar, “A Generalized Software Reliability Growth Model with different severity of faults” International Journal of Applied Studies, Vol. 3 Issue 11, 2014.

[8] D N Goswami, Anshu Chaturvedi and Mohammad Altaf Dar, “Software Reliability Growth Model with varying-Time fault removal efficiency as well as with fault Introduction” International Journal of Science and Research, Vol. 4 Issue 2, 2015.

[9] Mohammad Altaf Dar, D N Goswami and Anshu Chaturvedi, "Generalized Framework with Different Severity of Faults for Modelling Software Reliability Growth during Testing", International Journal of Advanced Research in Computer Science & Technology, Vol. 3, Issue 1, 2015.

[10] Mohammad Altaf Dar, D N Goswami and Anshu Chaturvedi, "Testing effort dependent Software Reliability Growth Model with dynamic faults for debugging process", International Journal of Computer Applications, Vol. 113, No. 11, 2015.

[11] Mohammad Altaf Dar, Showkat Ahmad Teeli and Fayaz Ahmad Bhat, Framework For Modelling Software Reliability Growth For Error detection With Dynamic Faults", International Journal of Advanced Scientific Research and Management, Volume 3 Issue 9, Sept 2018

[12] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, 2007.

[13] A Cloud Securty Alliance Collaborative research, "Expanded Top Ten Big Data Security andPrivacy challenges" , April 2013.

[14] Privacy-Preserving Ciphertext Multi-Sharing Control for Big Data Storage Kaitai Liang, Willy Susilo, Senior Member, IEEE, and Joseph K. Liu 2015.

[15] Privacy Preservation in the Age of BigData :A SurveyJohn S. Davis II, Osonde A. Osoba