Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210946

Inference Control in a Diabetes Data Set Using a Java-Based Prototype of LDH Algorithm

Georgios FERETZAKIS[1], Konstantinos MITROPOULOS, Dimitris KALLES and Vassilios S. VERYKIOS
School of Science and Technology, Hellenic Open University, Patras 263 35, Greece; georgios.feretzakis@ac.eap.gr; kmitrop@otenet.gr; kalles@eap.gr; verykios@eap.gr

Abstract. Data sharing among different entities in the healthcare domain has become an increasingly common practice, where each entity would most likely want to prevent indirect data disclosure via inference channels. The Local Distortion Hiding (LDH) algorithm has been developed to protect sensitive decision tree (DT) rules, which are chosen not to be disclosed when DT construction techniques are applied to the data. This article presents eight experiments using a Java-based prototype that implements the LDH algorithm in a diabetes data set. Our experiments test the ability of the LDH algorithm in two ways, firstly in inference control and secondly in maintaining the structure and the performance metrics of the resulting DT. Our experiments on hiding eight terminal nodes in a diabetes data set using a Java-based prototype that implements the LDH algorithm, yield satisfactory results.

Keywords. Inference control; data security; privacy-preserving; machine learning

1. Introduction and Background

The healthcare sector is being digitally transformed by technological advances in medical information systems, electronic medical records, wearables, and mobile devices. The increase in the amount of global healthcare data and the advancements in the machine learning (ML) and data analytics field allow researchers and clinicians to extract and visualize large-scale medical data in a new spectrum [1]. The Internet facilitates the transfer and the exchange of these data, as well as the delivery of healthcare services and applications, linking this way successfully patients and healthcare providers. While such ecosystems promise a future for widely accessible and more innovative healthcare, the privacy of patients, physicians, nurses, and health care professionals is today more than ever of concern [2]. Data privacy is a critical issue in health informatics, particularly when analyzing datasets collected from various sources, such as health care providers, insurance companies, pharmaceutical companies, and research institutions. Data sharing among different entities in the healthcare domain has become an increasingly common practice, where each entity would most likely want to prevent indirect data disclosure via inference channels. The extraction of knowledge from patients' personal data for research purposes should be made with safety and absolute privacy. Privacy-preserving

[1] Corresponding Author, Georgios FERETZAKIS, PhD; E-mail: georgios.feretzakis@ac.eap.gr

data mining [3] is a research field designed to resolve confidentiality issues arising from data mining. The inference problem in databases occurs when sensitive information can be disclosed via inference channels from non-sensitive data and metadata [4].

1.1. A related to the inference problem scenario

A scenario that could present an inference problem is the following. Let us consider Mary Johnson, who works for a big company. Once a year, the employee association, which is primarily concerned with the welfare and recreational activities, organizes a research study in which all the employees are asked to complete questionnaires n a voluntary basis about their eating habits, their lifestyle, and their health status. Think about the case where the association offers to some diabetic employees coupons for discounted products, given that these individuals have chosen to hide while filling in the questionnaires the fact that they have diabetes. Last year, for the first time, the company's administration asked the employee association for this data set to analyze it, for the benefit of the employees, by using ML techniques to create inference rules that will help accommodate the employees' needs better. The association wonders whether it should provide the data set to the company's administration team mainly because of the indirect disclosure of sensitive data of certain employees through inference channels. One of the paths (rules) of the DT that was deduced from this data set matches Mary's profile, who recently was diagnosed with diabetes, inferring in this way that Mary must be a diabetic. This conclusion was made even if Mary did not provide this information through the questionnaires, mainly because still other individuals who provided this information for themselves share the same habits with Mary. Deletion of the data concerning all the individuals that match the user profile of Mary, which was used to construct the DT would be an insufficient fix to the inference problem since important information that may be needed for other reasons will be lost accidentally. The best approach to address this issue would be to apply an inference control mechanism to hide some specific nodes of the DT without disturbing the overall structure of the original tree and losing the intrinsic value of the remaining rules.

1.2. Background

In articles [5-8], the authors proposed a series of strategies that would effectively protect against the disclosure of the sensitive classification rules. The LDH algorithm [9] was developed on the basis of the concept of preserving sensitive DT rules resulting from the use of data mining techniques. LDH algorithm minimally changes the initial dataset, resulting in the DT generated through hiding being syntactically close to the original one. The modified dataset generated by the LDH algorithm can be shared without any concern for disclosing the sensitive rule. In article [10], the authors presented a Java-based prototype that implements the LDH algorithm.

2. Applying LDH Algorithm for Inference Control Purposes

In our experiments, we used various performance metrics to compare the efficiency of the deduced DT with the original one. One of the most popular algorithms for rule-based classification, the C4.5 algorithm [11], uses the gain ratio as the splitting criterion. In every iteration, the attribute with the highest gain ratio is chosen as the splitting attribute.

Therefore, if we want to suppress a specific attribute test at a node, it would be a reasonable heuristic to change the values of the instances that would enter that node. By this change, the resulting gain ratio (due to that attribute) will be decreased and be equal to zero, where possible. The LDH algorithm locates the parent node of the leaf to be hidden and ensures that the attribute tested at that node will not generate a splitting, which would allow that leaf to re-emerge. A schematic diagram of the LDH algorithm's workflow is shown in Figure 1.

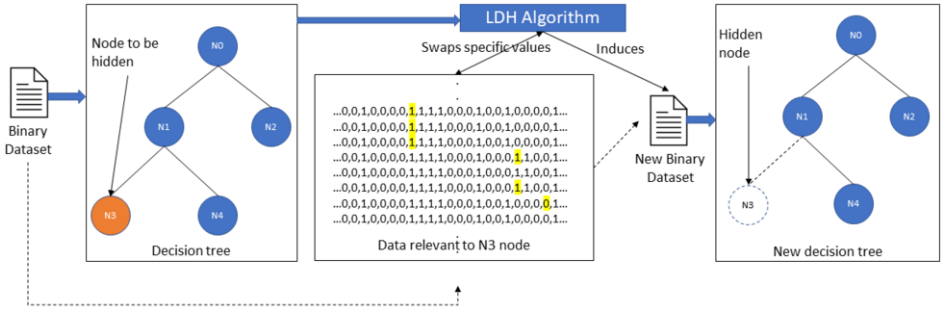

Figure 1. A schematic diagram of the LDH algorithm's workflow

The Java-based prototype [10] uses the Weka library to read an Attribute-Relation File Format (ARFF) data set and visualize the J48 DT with the provided interactive tree visualizer class. Afterward, the user may select an attribute represented with a node in the tree visualizer window, and subsequently, the prototype implements the LDH algorithm. The output of the software is the modified data set and the resulting DT.

3. LDH Implementation on a Diabetes Dataset

This section shows an example regarding an early-stage diabetes risk prediction data set [12] from the UC Irvine Machine Learning Repository [13]. This data set contains reports of 520 persons who have recently become diabetic or are still nondiabetic but have few or many symptoms. The data set includes fifteen binary attributes and one numerical (Age). We modified the original data set by removing Age's numerical variable since the LDH algorithm works only with binary features. The binary variables are Sex, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, and Obesity. The corresponding values for the above binary variables are Yes or No, whereas the class variable represents whether the patient is having a risk of diabetes (positive) or not (negative). We chose for our experiments to use the WEKA [14] framework, an ML software written in Java, and, more specifically, the *J48* classifier, the implementation of the C4.5 algorithm. We performed eight experiments using the Java-based prototype of the LDH algorithm [10] to evaluate the impact of each hiding regarding the DT's performance metrics. Each hiding was applied to every one of the eight terminal nodes of the original DT. Since the DTs mentioned above are too big to fit into a single page, they are available on the website [15]. All dataset files (.arff), before and after hiding, was applied, and the corresponding outputs are also available on the website [15]. Each of the eight terminal nodes was successfully hidden in all experiments, and the corresponding DTs were syntactically close to the original DT. The main performance

metrics of the original DT and the DTs induced from the modified data sets are presented in Table 1.

Table 1. Main performance metrics of the DTs induced from the original and the modified datasets.

	Original	Modified
Kappa statistic	0.9878	[0.9756-0.9878]
Mean absolute error	0.0084	[0.0084-0.0165]
Root-mean-squared error	0.0648	[0.0648-0.0907]
Relative absolute error	1.772%	[1.772%-3.478%]
Root relative squared error	13.314%	[13.314%-18.652%]

4. Conclusion

In this article, several experiments on inference control in a diabetes data set are presented where eight terminal nodes are hidden using a Java-based prototype that implements the LDH algorithm, all of which yielded satisfactory results.

References

[1] Nazir S, Khan S, Khan HU, Ali S, Garcia-Magarino I, et al. A Comprehensive Analysis of Healthcare Big Data Management, Analytics and Scientific Programming. IEEE Access. 2020;8:95714–33.
[2] Iyengar A, Kundu A, Pallis G. Healthcare Informatics and Privacy. IEEE Internet Comp. 2018;22:29–31.
[3] Verykios VS, Bertino E, I. Fovino I, Provenza L, Saygin Y, Theodoridis Y. State-of-the-art in privacy-preserving data mining. ACM SIGMOD Record 2004;33:50.
[4] Farkas C, Jajodia S. 2002. The inference problem: a survey. SIGKDD Explor. Newsl. 2002;4:6–11. doi:https://doi.org/10.1145/772862.772864
[5] Kalles D, Verykios VS, Feretzakis G, Papagelis A. Data set operations to hide decision tree rules. In Proceedings of the Twenty-second European Conference on Artificial Intelligence, Hague, The Netherlands, 29 August–2 September 2016.
[6] Kalles D, Verykios VS, Feretzakis G, Papagelis A, Data set operations to hide decision tree rules. In Proceedings of the 1St International Workshop on AI for Privacy and Security—Praise '16, Hague, The Netherlands, 29–30 August 2016.
[7] Feretzakis G, Kalles D, Verykios VS. On Using Linear Diophantine Equations for Efficient Hiding of Decision Tree Rules. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence—SETN '18, Patras, Greece, 9–12 July 2018. doi:https://doi.org/10.1145/3200947.3201030
[8] Feretzakis G, Kalles D, Verykios VS. On Using Linear Diophantine Equations for in-Parallel Hiding of Decision Tree Rules. Entropy 2019;21(1):66. doi: https://doi.org/10.3390/e21010066
[9] Feretzakis G, Kalles D, Verykios VS. Using Minimum Local Distortion to Hide Decision Tree Rules. Entropy 2019;21:334, doi: https://doi.org/10.3390/e21040334
[10] Feretzakis G, Mitropoulos K, Kalles D, Verykios VS. Local Distortion Hiding (LDH) Algorithm: a Java-based prototype. In 11th Hellenic Conference on Artificial Intelligence (SETN 2020). Association for Computing Machinery, New York, NY, USA, 144–149. doi:https://doi.org/10.1145/3411408.3411419
[11] Quinlan JR, C4.5. Programs for Machine Learning. Morgan Kaufmann: CA, USA, 1993.
[12] Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. Computer Vision and Machine Intelligence in Medical Image Analysis Advances in Intelligent Systems and Computing. 2019:113–25.
[13] Dua, D, Graff C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] 2019. Irvine, CA: University of California, School of Information and Computer Science.
[14] Frank E, Hall MA, and Witten IE. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
[15] Feretzakis G. Hiding Decision Tree Rules in a Diabetes Data Set using LDH Algorithm Java-based prototype. Available online: http://www.learningalgorithm.eu/datafiles_ICIMTH2021.html (accessed on 17 July 2021).