International Journal of Information Processing, 7(3), 68-82, 2013
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Two Stage Prediction Process with Gradient Descent Methods Aligning with the Data Privacy Preservation

S kumarasawamy[a], Srikanth P L[a], Manjula S H[a], Venugopal K R [a], L M Patnaik[b]

[a]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001 India, Contact: kumar.aruna@gmail.com.

[b]Honorary Professor, Indian Institute of Science, Bangalore.

Privacy preservation emphasize on authorization of data, which signifies that data should be accessed only by authorized users. Ensuring the privacy of data is considered as one of the challenging task in data management. The generalization of data with varying concept hierarchies seems to be interesting solution. This paper proposes two stage prediction processes on privacy preserved data. The privacy is preserved using generalization and betraying other communicating parties by disguising generalized data which adds another level of privacy. The generalization with betraying is performed in first stage to define the knowledge or hypothesis and which is further optimized using gradient descent method in second stage prediction for accurate prediction of data. The experiment is carried with both batch and stochastic gradient methods and it is shown that bulk operation performed by batch takes long time and more iterations than stochastic to give more accurate solution.

Keywords : Batch Gradient, Gradient Descent, RDF and Ontology, Stochastic Gradient.

1. INTRODUCTION

Data mining is a process of extracting the knowledge from large set of data. The knowledge extraction defines a model or rules for performing the accurate analysis on the future data. The initial data set is referred as training data, since it is used as reference for arriving at knowledge. The analysis performed on the data should not reveal the data as such and hence preserving the privacy of data becomes significant. Intuitively, differential privacy ensures that the system behaves essentially the same way, independent of whether any individual, or small group of individuals, opts in to or opts out of the database [1]. Generalization and suppression are two predominant techniques to achieve the privacy of data [2-3]. Privacy of data is very important in certain domains like hospital, analysis of psychological behaviour of patients [4]. The solution presented in this paper is to apply gradient descent methods on the privacy preserved data. The gradient descent is first order optimization algorithm to find local minimum of the function [5].

1.1. Motivation

Gradient descent is a widely used paradigm for solving many optimization problems. In machine learning or data mining, this optimization function corresponds to a decision model that is to be discovered [6]. Shuguo Han *et. al.,* has proposed the solution for application of gradient descent methods on the privacy preserved data. The data is either vertically or horizontally partitioned across communicating parties. The factor for partition is mutually synchronized between each other. In vertical partitioning every communicating party has same set of objects with varying attributes. In this scenario the prediction is first performed on unknown attributes before performing prediction for specific object. In horizontal partitioning every communicating party has disjoint set of records with same set of attributes. The prediction in this scenario is performed on unknown objects. To achieve required accuracy

stage prediction process. The identified maximum values are disguised by adding 10$ to the amount.

Alice and *Bob* performs the optimization of regression model obtained as a result of first stage prediction process. The optimization is achieved by applying gradient descent methods on the predicted output. In gradient descent the weight vector is used as a coefficient for prediction function as defined in Eq. (5) and the weight vector is optimized as per Eq. (6) during stochastic gradient and as per Eq. (7) during batch gradient descent. In stochastic gradient descent the optimization happens in less factor in each iteration as it update one element of weight vector at a time whereas in batch gradient the optimization happens with high factor.

As a result of which the batch gradient descent takes less number of iterations to predict output for high minimization factor/learning stoppage as shown in Figure 4, which indicates for high minimization factor/learning stoppage the batch gradient descent method takes more iterations until switching point of 0.5 minimization factor after which the behavior of batch and stochastic remains stagnant. The stochastic gradient takes less number of iterations for prediction after 0.5 *i.e.*, for lower values of minimization factor. Similarly, the execution time for batch gradient is high as number of iterations are more for prediction up to switching point as shown in Figure 5. After switching point the stochastic the batch gradient takes more time than stochastic for lower values of minimization factor. The experiment is carried out by varying learning stoppage from high value to low and behavior of stochastic and batch gradient descents are analyzed with respect to number of iterations and time required to perform prediction.

The DES algorithm is used as cryptographic algorithm to retain confidentiality of data exchanged between *Alice* and *Bob*. The experiment is simulated with raw sockets and implemented using JAVA. The *Bob* socket is used as server socket waiting for *Alice* to initiate communication. In First stage prediction the RDF location is exchanged securely and in second stage the regression model which is obtained as result from first stage prediction is exchanged securely between *Alice* and *Bob*.

9. CONCLUSIONS

In this paper a new approach is proposed to retain the privacy of data with the combination of generalization and bamboozling. The bamboozling is the process where the second level privacy is achieved by disguising the generalized information with certain factor to deceive the other communication parties. The complete generalization and bamboozling process happens as first stage prediction process to define the regression model which is used as input for second stage prediction for optimization and predict accurate result.

The experimental results shows the behaviour of batch and stochastic process with respect to the number of iterations and execution time required for prediction process. It is shown that batch gradient descent takes long time and more iterations for higher values of minimization factor than stochastic method. The solution can be enhanced by adding the suppression technique along with generalization and bamboozling processes to further strengthen the privacy of data.

REFERENCES

1. Microsoft Research. Database Privacy, *http://research.microsoft.com/en-us/projects/DatabasePrivacy*.
2. Latanya Sweeney. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression, *In International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5): 571-588, 2002.
3. Gabriel Ghinita, Member, IEEE, Panos Kalnis and Yufei Tao. Anonymous Publication of Sensitive Transactional Data, *In IEEE Transactions on Knowledge and Data Engineering*, 23(2): 161-174, Feb 2011.
4. Hopoper N, Saunders J and McHugh L. The Derived Generalization of Thought Suppression, *Learn Behav*, 38(2): 160-168, 2010.

5. Microsoft research. Database Privacy, *http:// research.microsoft.com/en-us/projects/ DatabasePrivacy*
6. Shuguo Han, Student Member, IEEE Computer Society, Wee Keong Ng, Member, IEEE Computer Society, Li Wan and Vincent C S Lee. Privacy-Preserving Gradient-Descent Methods, *IEEE Transactions on Software Engineering*, 22(6):884-899, June 2010.
7. Afshar P. Gradient Descent Optimisation for ILC-based Stochastic Distribution Control, *IEEE International Conference on Control and Automation(ICCA)*, 11341139, 2009.
8. Zhi Ding, Junqiang Hu and Dayou Qian. On Steepest Descent Adaptation: A Novel Batch Implementation of Blind Equalization Algorithms, In *Global Telecommunications Conference (GLOBECOM 2010)IEEE*, 1-6, 2010.
9. Gannot S. Iterative-batch and Sequential Algorithms for Single Microphone Speech Enhancement, In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, 2:1215-1218, 1997.
10. Gonzalez A. A Note on Conjugate Natural Gradient Training of Multilayer Perceptrons, In *International Joint Conference on Neural Networks (IJCNN '06)*, 887-891, 2006.
11. Ningning Jia, E Y Lam. Stochastic Gradient Descent for Robust Inverse Photomask Synthesis in Optical Lithography, In *17th IEEE International Conference on Image Processing*, 2010.
12. S Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds, In *IEEE Transactions on Automatic Control*, 58(9): 2217-2229, 2013.
13. D Beckett. RDF/XML Syntax Specification (Revised), *http:// www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/*, 1994.
14. John Hebeler, Matthew fisher, Ryan Blac, Andrew perez-lopez. Semantic-Web Programming, *Third Edition, Wiley India pvt.ltd*, 2009.
15. Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann and Ian Horrocks. The Semantic Web: The Roles of XML and RDF, *IEEE Internet Computing*, pages 63-73, 2000.
16. Kanishka Bhaduri, Mark D Stefanski and Ashok N Srivastava. Privacy Preserving Outlier Detection through Random Nonlinear Data Distortion, *IEEE Transactions on Systems, Man and Cybernetics*, 41(1):260-272, 2011.
17. Benjamin C. M. Fung, Member, IEEE, Thomas Trojer, Patrick C. K. Hung, Member,IEEE, Li Xiong, Khalil Al-Hussaeni and Rachida Dssouli. Service-Oriented Architecture for High-Dimensional Private Data Mashup, *In IEEE Transactions on Services Computing*, 5(3):373-386, 2012.
18. Jung Yeon Hwang, Sokjoon Lee, Byung-Ho Chung, Hyun Sook Cho and DaeHun Nyang. Short Group Signatures with Controllable Linkability, In *Workshop on Lightweight Security and Privacy: Devices, Protocols and Applications*, 44-52, March 2011.
19. Zhu Yu-quan, Tang Yang Chen Geng. A Privacy Preserving Algorithm for Mining Distributed Association Rules, *In International Conference on Computer and Management (CAMAN)*, 1-4, May 2011.
20. Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi. Privacy-Preserving Updates to Anonymous and Confidential Databases, *In IEEE Transactions on Dependable and Secure Computing*, 8(4):578-587, July-August 2011.

Kumaraswamy S is currently working as an Assistant Professor in the Department of Computer Science and Engineering, KNS Institute of Technology, Bangalore, India. He obtained his Bachelor of Engineering from SiddaGanga Institute of Technology, Tumkur, Bangalore University, Bangalore. He is presently pursuing his Ph.D programme in the area of Privacy Management in Databases in Bangalore University. His research interest is in the area of Data Mining, Web Mining and Semantic Web.

Srikanth P L received his Master's degree from the Department Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. His research interest is in the area of Web Technology, Semantic Web and Cloud Computing.

S H Manjula is currently the Chairman, Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. She obtained her Bachelor of Engineering and Masters Degree in Computer Science and Engineering from University Visvesvaraya College of Engineering. She was awarded Ph.D. in Computer Science from Dr. MGR University, Chennai. Her research interests are in the field of Wireless Sensor Networks and Data Mining.

K R Venugopal is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored 39 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems *etc.*. During his three decades of service at UVCE he has over 400 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

L M Patnaik is currently Honorary Professor, Indian Institute of Science, Bangalore, India. He was a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India and was a Professor since 1986 with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.