International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Privacy-Preserving Data Mining using Random Decision Tree

Komal N. Chouragade[1], Trupti H. Gurav[2]

Pune University, SKNCOE, Pune, Maharashtra, India

Abstract: ***Distributed data is widely spread in advanced information driven applications. With various sources of data, main problem is to decide how to collaborate adequately crosswise over collaborate limits while maximizing the utility of gathered information. Since utilizing local data provides suboptimal utility, methods for privacy-preserving collaborative knowledge detection must be produced. Past cryptography-based work privacy-preserving data mining is still excessively slower to be powerful for huge scale data sets to handle today's huge data problem. Past work on Random Decision Trees (RDT) demonstrates that it is conceivable to produce proportionate and exact models with much more modest cost. The fact can be utilized fully that RDTs can characteristically fit into a parallel and completely distributed framework, and generate protocols to develop privacy-preserving RDTs that allow general and effective distributed privacy-preserving knowledge discovery.***

Keywords: Distributed data, Privacy-Preserving Data Mining, Random Decision Trees, Knowledge Detection, Protocol.

1. Introduction

Developing and applying any data mining model supposes that the basic data is available. Frequently, this is not practical. Confidentiality and security issues delimit the sharing or centralization of data. Privacy-preserving data mining is emerging out as an efficient technique to tackle this issue. Distributed solutions are presented which preserves privacy while even now empowering data mining. In any case, while perturbation based solutions do not provide full privacy, cryptographic solutions are not efficient and infeasible to empower substantial scale examination to face the time of huge data.

A technique can be implemented to tackle this issue which will use both randomization and cryptographic methods to give enhanced proficiency and security to a number of decision tree-based learning tasks. This technique can provide an order of magnitude enhancement in efficiency over previous techniques. It will also provide more security. This is an efficient technique to tackle the issue of privacy-preserving data mining in big data challenge.

1.1 Random Decision Tree

While the utilization of RDTs may appear unreasonable, there are numerous advantages as far as performance and accuracy that are picked up by utilizing this system versus conventional algorithms. Fan et al. [1] find that for classification, utilization of a random model can match, as far as solutions, other inductive learning models in discovering an optimal theory.

In the meantime, RDT performs better than different models with respect to computation speed, because of the characteristics of random partitioning utilized as a part of tree development. The RDTs algorithm constructs multiple (or m) iso-depth RDTs. One essential aspect of RDTs is that the structure of a random tree is built totally independent of the training data. The RDT algorithm can be divided into two steps, those are training and classification. The training step comprises of building the trees (BuildTreeStructure) and populating the nodes with training example data (UpdateStatistics).

It is considered that the quantity of attributes is known depended on the training data set. The depth of every tree is selected based on a heuristic—Fan et al. [1] demonstrate that the most diversity is accomplished, when the depth of the tree is equivalent to a half of the total number of features present in the data, preserving the benefits of random modeling. The procedure for generating a tree is as per the following.

1. Begin with a list of features that is attributes from the data set. Create a tree by randomly selecting one of the features without utilizing any training data. The tree halts growing as height limit is reached.
2. Utilize the training data to upgrade the statistics of every node. Note that just the leaf nodes require storing the number of illustrations of diverse classes that are classified through the nodes in the tree.

The training data is examined once to update the statistics in different random trees. At the point when classifying another example x, the probability outputs from numerous trees are averaged to calculate the a posteriori probability.

1.2 Horizontally Partitioned Data

At the point when data is horizontally partitioned, parties gather data for distinctive entities, yet have data for the each of the attributes. We now require to evaluate how the RDTs can be developed and how classification is obtained. As all the parties share the schema, a simple solution is for all parties to separately make some random trees.

Altogether these will create the ensemble of random trees. Besides that, every party can freely make the structure of the tree. All parties must co-operatively and safely calculate the parameters that are estimations of every leaf node, over the universal data set. Dissimilar to the basic RDT technique, there is no compelling reason to keep the class distribution at

every non-leaf node this data is just needed at the leaf nodes. Presently, there are two conceivable outcomes:

1) Every member is known of the structure of the tree.
2) Every member is unknown of the structure of the tree.

Inside first possibility, there are three further possibilities:

a. All parties will be known of the global class distribution vector for every leaf node.
b. Only the party owning the tree is known of the global class distribution vector for every leaf node.
c. No party is known of the global class distribution vector for every leaf node.

Inside second possibility, there are two further possibilities:

a. The tree owning party is known of the values for every leaf node.
b. No party is known of the values for every leaf node.

It is clear that case (2) is more complex than case (1), as all the parties are unknown of the structure of the tree. This creates an issue as other parties can no more calculate the local leaf values. Each party needs to cooperate with the tree owner somehow to calculate the leaf node values.

Then again, it does not make any sense. Initially, as everyone is known of the schema, and the tree structure is random, anybody could think of that specific structure. To be sure, it would really be ideal to remove random structures that may be not acceptable to a few parties because of privacy concerns. Also, regardless of the possibility that the structure is unknown, each classification of another example can uncover some knowledge of tree.

1.3 Vertically Partitioned Data

With vertically partitioned data, data for same set of entities is collected by all parties. Notwithstanding, every party gathers data for a different set of attributes. Presently the party cannot separately make even the structure of a random tree, unless they share the attribute data between them. Subsequently, there are two possible outcomes:

1. All parties share fundamental attribute data that is metadata. Presently they can freely generate random.
2. There is no sharing of data. Presently, the parties require to work together to generate the random trees. These trees could themselves exist in a distributed structure.

Not at all like the horizontal partitioning case, the structure of the tree does uncover conceivably sensitive data, since the parties do not comprehend what are the attributes possessed by other parties.

2. Literature Review

In the study by R. Agrawal and R. Srikant [2] considered the technical achievability of realizing privacy-preserving data mining. The fundamental reason was that the sensitive values in a user's record will be perturbed utilizing a randomizing function so they cannot be evaluated with efficient accuracy. Randomization is possible utilizing Gaussian or Uniform perturbations. The problem they attained was whether, given a substantial number of users who do this perturbation, would we be able to still build sufficiently exact predictive models.

For the particular instance of decision-tree classification, they discovered two powerful algorithms, ByClass and Local. The algorithms depend on a Bayesian methodology for adjusting perturbed distributions. They underline that they rebuild distributions, not individual records, therefore preserving privacy of individual records. In actuality, if the user perturbs a sensitive value once and regularly give back perturbed value, the assessment of the true value cannot be enhanced by progressive queries. They found in their observational assessment that:

a) ByClass and Local are both successful in correcting for the impacts of perturbation. At 25% and 50% privacy levels, the precision numbers are close to those on the original information. Indeed at 100% privacy, the algorithms were inside 5% to 15% of the original precision. Review that if privacy were to be measured with 95% certainty, 100% privacy implies that the true value cannot be assessed any closer than an interval of width which is the whole range for the relating attribute. They accept that a little drop in precision a desirable trade-off for privacy in number of scenario.
b) Local performed slightly better than ByClass, however obliged significantly more computation. Examination of what attributes may make Local a winner over ByClass is an open issue.
c) For the same privacy level, Uniform perturbation performed considerably poorer than Gaussian before correcting for randomization, however a little poorer after correcting for randomization. Subsequently the decision between applying the Gaussian or Uniform distributions to preserve privacy ought to be focused around different considerations: Gaussian gives more privacy at higher confidence thresholds, yet Uniform may be less demanding to clarify to users.

D. Agrawal and C.C. Aggarwal [4] examined the configuration and quantification of privacy-preserving data mining algorithms. They proposed an expectation-maximization algorithm which provably unites to the most maximum probability evaluation of the original distribution. Consequently, the algorithm gives a robust evaluation of the original distribution. They established the frameworks for quantification of privacy gain and information-loss in a hypothetically exact and technique independent way. They qualified the relative adequacy of diverse perturbing distributions utilizing these metrics. Their experiments additionally exhibit that when the data is extensive then the expectation maximization algorithm can rebuild the data distribution with nearly zero data loss.

Protecting privacy in data mining exercises is an imperative issue in numerous applications. Randomization - based procedures are liable to assume a vital part in this domain. Notwithstanding, H. Kargupta et. al [5] demonstrated a few of the difficulties that these systems confront in preserving the data privacy. It demonstrated that under specific

conditions it is moderately simple to break the privacy security offered by the random perturbation based methods. It gave widespread exploratory results with diverse sorts of data and demonstrated that this is truly a concern that must be attained. Notwithstanding to raise this concern they presented a random-matrix based data filtering methods that may discover more extensive application in generating another point of view for creating better privacy-preserving data mining algorithms.

The option methodology to ensuring privacy of distributed sources utilizing cryptographic methods was initially applied in the domain of data mining for the development of decision trees by Lindell and Pinkas [3]. This work comes under the structure of secure multiparty computation [6], attaining "flawless" privacy, i.e., nothing is learned that could not be derived from one data and the ensuing tree. The key understanding was to trade off computation and communication cost for precision, enhancing productivity over the generic secure multiparty computation strategy. In any case, the presented solution is still excessively wasteful for practical use. There has been work in distributed development of decision trees on vertically partitioned data.

Wang et al. propose an answer focused around passing the transaction identifiers between sites [7]; while this does not uncover particular attribute values, parties realize which exchanges take after which way down the tree, empowering one site to say "these two particulars have the same attribute values." Du and Zhan [8] do propose a method for building a privacy-preserving decision tree classifier for vertically partitioned data. Their system is constrained to two parties, accepts that both parties have the class attribute, and is not actualized. Vaidya et al. [9] broaden this to the multi-party case, likewise unwinding the supposition that the class attribute must be known to each party. This methodology has been actualized; however the exploratory results propose that for substantial data, the computational complexity is high.

Additional classification methods incorporate protocols to construct a Naive Bayes classifier [10], while Wright and Yang [11] present a similar protocol for learning in the Bayesian network framework. These works make trade-offs in the middle of effectiveness and data disclosure; however all keep provable limits on disclosure.

Jagannathan et al. [12] present approaches to build a differentially private RDT classifier from a concentrated data set. Since the data is distributed, it cannot be utilized straightforwardly. There has additionally been work to attain association rules in horizontally partitioned data [13] [14], EM Clustering in on horizontally partitioned data [15], clustering in vertically partitioned data [16], [17], association rules in vertically partitioned data [18], [19], and generalized methodologies to lessening the number of "online" parties [20]. The technique for demonstrating the accuracy of the algorithm originates from secure multiparty computation [6] [21] [22]. As of late, there has been a restored enthusiasm toward this field, a great discourse can be found in [23]. Presently, assembling these into productive privacy-saving data mining algorithms, and demonstrating them secure, is a difficult task.

3. Proposed System

A technique is presented to securely build Random Decision Trees (RDTs) for both horizontally and vertically partitioned data sets. The presented protocols are implemented and computation and communication cost is examined, and security. Additionally the performance of the presented protocols is compared with the current ID3-based protocols. RDTs can give great security with high productivity.

4. Conclusion

General and productive distributed privacy preserving knowledge discovery is really achievable. The security and privacy suggestions are considered when managing distributed data that is partitioned either on horizontally or vertically across multiple sites, and the difficulties of obtaining data mining tasks on such data. Since RDTs can be utilized to create identical, exact and off and better models with much less cost, distributed privacy-protecting RDTs is presented. This methodology powers the way that randomness in structure can give solid privacy with low computation.

References

[1] W. Fan, H. Wang, P.S. Yu, and S. Ma, "Is Random Model Better? On Its Accuracy and Efficiency," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), 2003.
[2] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data, pp. 439-450, May 2000.
[3] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.
[4] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems, pp. 247-255, May 2001.
[5] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE Int'l Conf. Data Mining (ICDM '03), Nov. 2003.
[6] O. Goldreich, "General Cryptographic Protocols," The Foundations of Cryptography, vol. 2, pp. 599-764, Cambridge Univ. Press, 2004.
[7] K. Wang, Y. Xu, R. She, and P.S. Yu, "Classification Spanning Private Databases," Proc. 21st Nat'l Conf. Artificial Intelligence, pp. 293-298, 2006.
[8] W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," Proc. IEEE Int'l Conf. Data Mining Workshop on Privacy, Security, and Data Mining, pp. 1-8, Dec. 2002.
[9] J. Vaidya, C. Clifton, M. Kantarcioglu, and A.S. Patterson, "Privacy-Preserving Decision Trees over Vertically Partitioned Data," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 3, pp. 1-27, 2008.
[10] J. Vaidya, M. Kantarcioglu, and C. Clifton, "Privacy Preserving Naive Bayes Classification," Int'l J. Very

Large Data Bases, vol. 17, no. 4, pp. 879-898, July 2008.
[11] R. Wright and Z. Yang, "Privacy-Preserving Bayesian Network Structure Computation on Distributed Heterogeneous," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. 2004.
[12] G. Jagannathan, K. Pillaipakkamnatt, and R.N. Wright, "A Practical Differentially Private Random Decision Tree Classifier," Proc. IEEE Int'l Conf. Data Mining Workshops (ICDMW '09), pp. 114-121, 2009.
[13] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.
[14] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," Proc. ACM SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery (DMKD '02), pp. 24-31, June 2002.
[15] X. Lin, C. Clifton, and M. Zhu, "Privacy Preserving Clustering with Distributed EM Mixture Modeling," J. Knowledge and Information Systems, vol. 8, no. 1, pp. 68-81, July 2005.
[16] J. Vaidya and C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 206-215, Aug. 2003.
[17] G. Jagannathan and R.N. Wright, "Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 593-599, Aug. 2005.
[18] J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 639-644, July 2002.
[19] J. Vaidya and C. Clifton, "Secure Set Intersection Cardinality with Application to Association Rule Mining," J. Computer Security, vol. 13, no. 4, pp. 593-622, Nov. 2005.
[20] M. Kantarcioglu and J. Vaidya, "An Architecture for Privacy- Preserving Mining of Client Information," Proc. IEEE Int'l Conf. Data Mining Workshop on Privacy, Security, and Data Mining, pp. 37-42, Dec. 2002.
[21] A.C. Yao, "How to Generate and Exchange Secrets," Proc. 27th IEEE Symp. Foundations of Computer Science, pp. 162-167, 1986.
[22] O. Goldreich, S. Micali, and A. Wigderson, "How to Play Any Mental Game—A Completeness Theorem for Protocols with Honest Majority," Proc. 19th ACM Symp. Theory Computing, pp. 218-229, 1987.
[23] W. Du and M.J. Atallah, "Secure Multi-Party Computation Problems and Their Applications: A Review and Open Problems," Proc. New Security Paradigms Workshop (NSPW '01), pp. 11-20, Sept. 2001.