Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Privacy Preserving Data Mining in a Distributed Environment: An Experimental Avenue

[1]Mona Shah[*], [2]Dr. Hiren D. Joshi
[1]Research Scholar, RK University, Rajkot, Assistant Professor, JG College of Computer Applications, Ahmedabad, India
[2]Associate Professor and Director (I/C) School of Computer Science, Dr. Baba Saheb Ambedkar Open University, (BAOU), Ahmedabad, India

Abstract— Data mining is a crucial and innate branch of research applicable in every aspect and sphere. The core and scope of this subject receives sharp supplementary focus when the element of privacy is added to it. Privacy preserving algorithms need to be addressed with a layer of maintaining secrecy and still achieving accuracy. Accomplishing speed, cost optimization would be an additional benefit. Apparently this would impose restrictions on the solutions largely available for customary mining purpose or at times design entirely new ones. In this work, an attempt is made to address one such privacy preserving data mining problem where multiple parties are involved in a distributed fashion. Each of them has a data set having the same nature and they are contributing to the problem of classification for combine mining. The end result enfolds a smart and simple classifier and a multiparty model which offer credible performance. The entire solution has been tested in a simulated environment for multiple machines and gives adequately satisfactory results.

Keywords— Privacy preserving, data mining, distributed environment, algorithm, classification, Naive Bayes.

I. INTRODUCTION

The practice of data mining inherently existed in human, ages ago inadvertently, before it was discovered and declared officially as a multidisciplinary science. This branch not only includes conventional methods containing statistics but incorporates methods pertaining to artificial intelligence, machine learning, databases and networks. The classic definition of data mining describes it as extraction of non-trivial, implicit, previously unknown and potentially useful information or pattern of data from large databases [1]. When we store any information, we have a just reason to retrieve it. Data mining is implemented for prediction, pattern analysis, decision support, managerial support, risk analysis and many others. Such data analysis either could be done on an individual basis or for a group of individuals. When it is done collectively for a group, there may be reasons for concealing the data submitted for mining purpose. The reason for this confidentiality could be the nature of data which may be personal information or either it could threat to competitive business. Privacy preserving data mining is providing solutions to data mining problems on confidential data which may be given by several parties. This paper is organized into 4 sections: 1) Introduction 2) Problem Addressed and Algorithm 3) Experiment and Results 4) Conclusion

Most of the business requires data mining at different levels and different stages at recurrent intervals. In few situations there may arise a need for collective mining of data for people involved in the same business or profession. This will widen the possible scope of the nature of data under study and this expanded horizon will cover more specially the rare occurrences and all possible combinations. It would lead to a more near to exact sculpt for the pooled instances. The final result needs to maintain hiding of the data contributed by each business. The possible and evident reasons are the pooling of business or else the sensitive nature of information including medical, defence and/or personal information in general categories. Such data can be individual information like bank account numbers, passport number, social security details, criminal history, medical health information, defence secrets, credit card numbers and a few more to add [2]. This inbuilt nature of the problem of privacy preserving itself draws boundaries to the possible ways of finding solution.

A number of solutions have been proposed and implemented earlier although each of them comes with a trade off and limitations relevant to the nature of the problem. The solution given in [3] proposes a solution where the model suggests mining the data locally and sending it to the next party encrypted. This solution works well for two parties but not for large databases. One solution suggests [4] that there will be miner and calculator both. None have the actual database and this works well in horizontally and vertically partitions both. The encryption is done at miner and calculator level both. To quote one more solution [5], this uses samples selection and matrix decomposition method. Singular value decomposition method is used to distort the original data. A transition probability matrix is made using Naive Bayes classifier.

In the problem addressed, the multiple parties are connected in a distributed system. The data is collected from each node and a joint model is generated. The latter model generated is used by all to classify their data individually.

II. PROBLEM ADDRESSED AND ALGORTIHM

Privacy preserving data mining (PPDM) under this work addresses the solution of collective mining of data from parties, who are connected in a distributed environment. Here, we have assigned one node who does not participate in mining as the intermediate or middle node, which takes care of every exchange that happens among the nodes/parties. This middle node is the unknown party. So, by this it is implied that any node that participates in the mining process by contributing its data only connects to the intermediate node. Also, the data set instances at each node have the same schema of data, same number and nature of attributes. Logically, it can be contemplated as a horizontal partition of a data set. The use of randomisation is done at two levels: primarily at the selection of the node(s), which will furnish the data set and secondarily at the selection of miner node. The nodes may play dual role viz: of participant essentially and of miner partially. This will augment the level of concealment as no node is aware of its stage of selection or assignment as miner prior to the starting of the process. Not only that the data will be sent only from the intermediate node, so the location identity viz : from where the data is being sent, appears only from the same intermediary node every time. Hence, the node acting as miner will never be able to judge the location of the data received from. The maximum amount of release of information would be that the received data belongs to some node. The initial masking of data has been done at each node for its data which would remove the instance identifiable parameters from the data set. The wrapper selection algorithm of Weka has been used, which is classifier specific for each data set instances in order to identify the parameters influencing the most to the different training data sets. With the process beginning, two nodes are randomly selected by the intermediate party and those two send in their data. The next move would be to select a miner randomly and send the training set received from two nodes to generate a model. The model is built using the evaluation method of holdout set. This model generated would then be send to the intermediary node.

When all such intermediate models arrive at the intermediary node, they are amalgamated to generate the final model. This classifier thus generated is sent back to all the nodes who contributed and they can envisage their instances. The accuracy at each node for the resultant model is also verified and the results are presented. The resultant model when tested multiple times, taps a significant level of acceptance each time.

The Algorithm:

Step 1: Obtain database files at the middle node.
Step 2: Generate model at miner node.
Step 3: Amalgamate the model(s) at middle mode.
Step 4: Repeat step from step 1 to step 3 till all nodes participating are considered.
Step 5: Send the resultant model to all participating nodes.

III. EXPERIMENT AND RESULTS

The experiment has been carried out on a system of 64 bit Intel i5 processor in a simulated environment of Oracle VM Virtual box (v. 4.3.28) with 6 virtual machines:5 participating machines and one middle node. Virtual box runs on a number of 32 bit and 64 bit host operating systems and is also referred as host hypervisor system. The host OS is Windows 8.1 and the guest OS is Centos Linux 7.Each node has been assigned a base memory of 1GB.Since the data is sensitive dealing with applicants/users of card, data sharing is stringently denied by the officials. Hence the use of synthetic data has been made. The synthetic data has been generated for five nodes each with 1000 instances at each node. The classification value for the data set is either yes or no. The data files consist of varied attributes counting to twenty six. The type of attributes comprises of nominal, categorical as well as numeric. The intermediary node is decided as the unknown party as per the model described in the work [6]. The intermediary node interacts with all five nodes. The participating nodes do not interact with each other. The entire system is programmed under a set of platform independent Java programs with a layer of shell scripting. Naive Bayes Classifier is used for the experiment.

The choice of Naive Bayes Classifier is a result of its simplicity, robustness and easiness to implement. It does not over fit the data. Naive Bayes performance remains optimal in case whether or not participating attributes are independent of each other **[7]**. The model can be modified with new training set without having to rebuild the model. The model is so small and simple that it is easy for anyone to learn. Naive Bayes can quickly adapt the changes in the data. It performs well with small data sets and even with missing data.

Once the final model is generated, it is passed to all the five nodes who had contributed in model building process. The final model is tested at each node. The system has been tested five times and each time, the combination of training files is different because of choice of randomness. The result comprises of two tables and one chart. The interpretation of the findings can be summarised as below.

1. Table I has the component accuracy, which shows the total number of correctly identified occurrences. The table shows the accuracy which is carried out individually at each node before building the model and after building the combined model. At each node, the accuracy shows improvement in the accuracy with the combined model.
2. Table II has five evaluation measures. TP Rate (TPR) also known as sensitivity signifies the proportion of "yes" correctly identified. Mathematically, it can be depicted as TP/ (TP+FN), where TP is true positive is correctly identified positive cases and FN represents incorrectly rejected instances **[08]**. FP Rate (FPR) is given as FP/ (FP+TN). FP is the number of instances incorrectly accepted negative instances. TN stands for True Negative which is the number of correctly accepted negative instances. Precision is the fraction of instances retrieved that are of importance to the instances retrieved. It is given as TP/ (TP+FP). Recall is the portion of data that are of

need and are successfully retrieved. It is same as TP rate. F-measure, a combine metric of precision and recall conveys the balance between the two. It is given as harmonic mean of precision and recall [9]. It is calculated as 2xprecisionx recall/ (precision+ recall). F- Measure is found to be best when it reaches value 1 and worst when at 0. From the table-II, it can be observed that in most of the cases, the F-Measure reaches to the exactness.

3. Figure-I is the comparison chart of accuracy observed in the data sets before applying combine mining and after applying combines mining. The X-Axis has the nodes and the Y-axis shows the accuracy in percentage observed before and after applying combined model respectively.

Table - I

Node	Before Accuracy %	After Accuracy %
Node1	85	85
Node2	85	93
Node3	85	90
Node4	89	92
Node5	78	85

Table – II

NODE	CLASS	TP Rate	FP Rate	Precision	Recall	F-Measure
Node1	YES	0.806	0.030	0.964	0.806	0.878
	NO	0.939	0.394	0.705	0.939	0.805
Node2	YES	0.903	0.016	0.982	0.903	0.941
	NO	0.974	0.158	0.860	0.974	0.914
Node3	YES	0.871	0.014	0.984	0.871	0.924
	NO	0.967	0.300	0.763	0.967	0.853
Node4	YES	0.879	0.000	1.000	0.879	0.935
	NO	1.000	0.235	0.810	1.000	0.895
Node5	YES	0.790	0.032	0.961	0.790	0.867
	NO	0.947	0.342	0.735	0.947	0.828

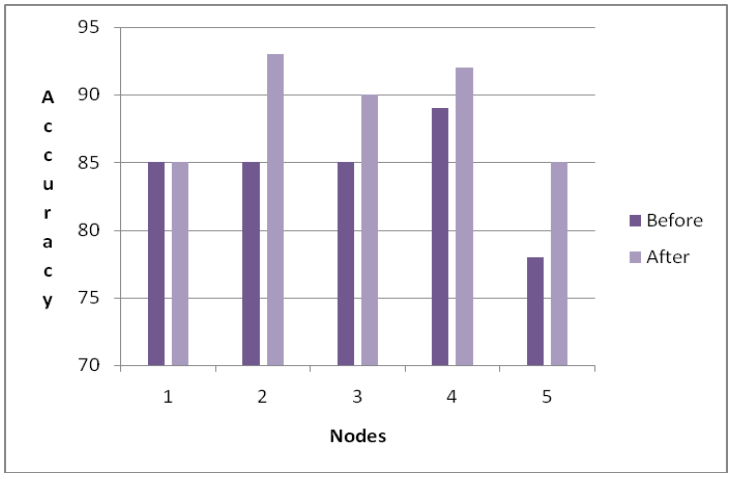

Figure – I

The final model also classifies the test data set. The value yes or no will determine whether the applicant's application should be approved or not based on the prior facts.

So, with the resultant miner model, we can complete the following task:

1: Classify all the new test data sets.

2. Test the model for the existing data sets.

Since at each stage the model has been built with a combination of training and testing the robustness of the model is stalwart enough for the business to sustain, flourish and expand.

IV. CONCLUSION

PPDM problems put forwards the challenge of mining task with keeping privacy. The current problem addresses a classification task among multiple parties in a distributed milieu with privacy preserving to categorize for the applicants of card. An experiment has been carried out using Naive Bayes classifier among five different connections. Privacy preserving mining and generating new classifier in the undertaken problem achieves acceptable result for a distributed system with multiple parties. Maintaining secrecy with accuracy is achieved with the proposed architecture and algorithm.

The scores that have been missed out in perfection can be attributed to external, unforeseen and immeasurable factors and one those affected by personal judgements.

REFERENCES

[1] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge Discovery in Databases: An Overview*", AI Magazine* Volume 13 Number 3 (1992)

[2] Shah Mona, Joshi Hiren D,, "Privacy Preserving Data Mining Techniques in a Distributed Environment", *International Journal of Computer Applications* (0975 – 8887) Volume 94 – No 6, May 2014

[3] Yang Z. and Wright R. N. Member IEEE, "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data". *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 18, NO. 9

[4] Gurevich A. and Gudes E., "Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation". 10th International Database Engineering and Applications Symposium (IDEAS'06), IEEE.

[5] Guang Li and Yadong W, "Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition", International Conference on Internet Computing and Information Services.

[6] Shab Mona, Joshi Hiren D, "Privacy Preserving Data Mining in a Shard Database: Architectural Aspect", ,SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – volume 2 issue 3 March 2015

[7] Harry Zhang, "The Optimality of Naive Bayes*", AAIG*, 2004

[8] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[9] https://en.wikipedia.org/wiki/Precision_and_recall

