ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

Impact of Encryption Techniques on Classification Algorithm for Privacy Preservation of Data

Jharna Chopra[1], Sampada Satav[2]

M.E. Scholar , CTA, SSGI, Bhilai, Chhattisgarh, India [1]

Asst.Prof, CSE, SSGI, Bhilai, Chhattisgarh, India[2]

Abstract:In this paper, the Naïve Bayesian and K-Nearest neighbour algorithms have been implemented for classification and AES, Triple DES and Rijndael on nine real-world datasets. The goal of the research is to evaluate the performance of the classification algorithms when the data set is encrypted using a variety of performance metrics: classification accuracy, precision, recall (sensitivity), specificity and lift charts/gain charts and to determine the impact of encryption on these algorithms. We found that aside from the obvious time penalty the implementation of an encryption algorithm to protect user privacy the performance of the classification algorithms remained the same in most of the datasets. However, the time penalties for encrypting the data before it could be used for classification varied greatly depending on the type of algorithm used to encrypt the data.

Keywords: Classification, data mining, statistical methods, logistic regression, regression trees, discriminant analysis.

I. INTRODUCTION

In the recent past, there has been an exponential increase in the amount of stored data. Managers and decision makers are faced with the problem of information overload. For example, in 1992, Frawley, Piatetsky-Shapiro and Matheus reported that the amount of data in the world doubles every twenty months. Cios, Pedrycz, and Swiniarski in 1998 reported that, Wal-Mart alone uploads twenty million point of sale (POS) transactions every day. Today we have far more information stored than we can handle. But as data volume increases, making meaningful decisions becomes increasingly difficult. To address these issues, researchers turned to a new research called Data Mining and Knowledge Discovery in Databases. In the past decades data mining methods have been widely used for the purpose of extracting knowledge from large data. Classification, a supervised method used to partition variables into several classes, represents the most widely used data mining method. But with this increasing volume of data comes the question of data privacy. How can we process this huge volume of data while keeping user privacy intact?

There have been several studies on comparing classification algorithms. However, most of these studies have been performed without taking into account the privacy issue. The theme of this paper is to classify various datasets using two of the most popular classification algorithms but the datasets will be encrypted to understand the pros and cons of enabling privacy preservation.

II. STATEMENT OF THE PROBLEM

An abundance of classification algorithms have been developed to solve data classification problems. Machine learning and data mining are among the most highly researched fields in today's world. However, the applications of the algorithms vary greatly with the scenario under consideration. A number of commercial tools are also available today which provide a wide range of classification techniques. No single algorithm, in all scenarios, has been demonstrated to be superior. Similarly, since a lot of the data being classified using these algorithms is personal to the user, it is also very important to consider which encryption algorithm should be used. "What is the impact of encryption on performance of data classification algorithms?" The primary focus of my research will be to evaluate the impact of encryption on the performance of two of the most popular classification algorithms using both statistical and machine learning methods on multiple datasets. An important aspect of my thesis is to use a variety of performance criteria to evaluate the learning methods. The performance criteria we have chosen to evaluate the algorithms are precision, recall

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

and specificity.The dataset chosen for the projectis the Newsgroup dataset.

III. EXPERIMENTAL PROCEDURE

The following classification algorithms have been selected within the scope of this paper. They are:

- K Nearest Neighbour and
- Naïve Bayesian

The following encryption algorithms have been selected within the scope of this paper. They are:

- AES,
- Triple DES and
- Rjindael

There are three phases to building this project : Building the software which implements all the above mentioned algorithms by using standard implementations, verifying the algorithms by running the algorithms on a sample dataset and checking the results and finally running the classification algorithms on datasets encrypted with the above mentioned encryption algorithms. The models were evaluated using the following evaluation methods:

- Precision,
- Recall/Sensitivity, and
- Specificity

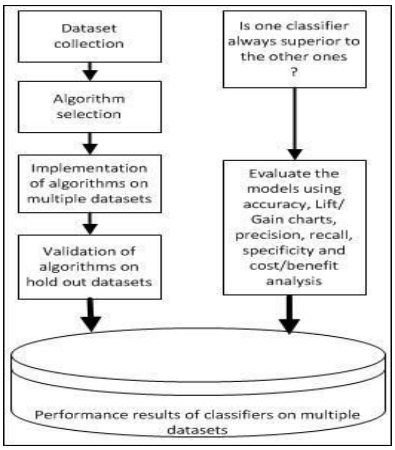

Fig.1 Proposed Methodology Framework

Fig 1 describes the proposed methodology for privacy preservation of data using various encryption algorithms.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

IV. RESULTS AND DISCUSSION

In this section the performance results of each algorithm will be discussed and the research question will be addressed. The performances of the selected algorithms were evaluated on the publicly available dataset.

A) Classification Accuracy

Table 1 shows, for each dataset, the estimated classification accuracy of the algorithms with and without encryption. As one can see from Table 1, the classification accuracy of the Naive Bayesian algorithm tends to be better while encryption is being used. The results also show that the effects of enabling encryption on the accuracy of the algorithms is minimal. AES and Rjindael show slightly better accuracy than TripleDES. The classifiers show comparable results even with encryption enabled.

Table 1

A comparison of the accuracy of Naive Bayes (NB) and K-Nearest Neighbour (KNN) classifiers without encryption and with AES, Triple DES and Rjindael on the given dataset.

Encryption Method	KNN	NB
None	84.27	86.94
AES	83.18	85.67
Triple DES	81.35	83.58
Rjindael	82.98	85.61

B) Recall, Precision and Specificity

Table 2 shows the confusion matrix for the neural networks (NN) classifier trained on the white wine dataset. We will use this tableto illustrate our evaluation techniques for recall, precision and specificity. The table cells represent the number counts in the test dataset. The columns represent the predicted class and the rows represent the actual class in the dataset. We can see from the table that the NN could not predict classes 8 and 9. For example, the number of samples with actual label 6 that were incorrectly predicted as 5 or 7 is 101 and 50 respectively. The *ATotal* column indicates the number of test samples whose actual label is specified by the row. For example, suppose we are interested in class 6. From the table, 558 samples were actually labeled 6: the cell shaded green is the number of true positives (TP). The cells shaded orange represent falsepositives (FP), the cells shaded yellow represent the false negatives (FN) and the cells shaded blue indicate true negatives (TN).The *PTotal* row indicates the number of test samples whose predicted label is specified by the column label. For example 720 samples had been predicted as 6. From Table 2, the TP=405 and the FP = 315 =3+10+135+136+30+1 see color coding. Therefore the precision for class 6 is:

$$P_6 = \frac{405}{(405+315)} = \frac{405}{(720)} = 0.5625$$

Here a total of 315 samples were incorrectly predicted as 6. The precision, recall and specificity for each class are

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

calculated. The overall precision, recall and specificity are computed as a weighted average. The results of the recall, precision and specificity are tabulated in different tables.

Table2

Confusion matrix for Neural Networks (NN) classifier trained on the white wine dataset. Note that the NN could not predict two of the classes (i.e. 8 and 9).

Class	3	4	5	6	7	*ATotal*
3	0	0	1	3	0	4
4	1	2	28	10	0	41
5	0	2	209	135	3	349
6	1	1	101	405	50	**558**
7	0	1	7	136	86	230
8	0	0	0	30	11	41
9	0	0	0	1	0	1
PTotal	2	6	346	**720**	150	1224

C) Performance by Dataset

Tables 3 below, show the statistics of the models for each problem (dataset). The C5.0 algorithm has the best accuracy for the adult, house, segment, white wine and the red wine datasets; naïve Bayes (NB) has the best accuracy for the NHANES, and cars datasets; neural networks (NN) has the best accuracy for the credit and the vehicle. Logistics regression (LR) tied with NB for the best accuracy for the NHANES dataset. CHAID, support vector machines (SVM), discriminant analysis (DA), QUEST, classification and regression trees (CART) never produced best accuracy result for any of the datasets. Overall classification accuracy alone does not distinguish between types of errors the classifier makes (i.e. False Positives versus False Negatives). For example two or more classifiers may exhibit the same accuracy but may behave differently on each category.

Table 3

Statistics for the Models by Problem; Accuracy, Recall, Precision and Specificity**.**

Model	Accuracy	Recall	Precision	Specificity
C5.0	66.01	0.386	0.457	0.924
NN	57.35	0.250	0.296	0.900
LR	54.9	0.261	0.524	0.897
CART	54.9	0.226	0.235	0.893
CHAID	53.68	0.233	0.230	0.894
QUEST	50.49	0.177	0.210	0.874
SVM	49.84	0.221	0.400	0.894
NB	47.47	0.365	0.406	0.915
DA	29.74	0.338	0.225	0.878

A - White Wine dataset

Model	Accuracy	Recall	Precision	Specificity
C5.0	97.4	0.975	0.976	0.996
CART	96.19	0.962	0.962	0.994
NN	96.01	0.961	0.957	0.993
CHAID	95.67	0.958	0.958	0.993
LR	93.59	0.935	0.932	0.989
SVM	92.72	0.929	0.930	0.988
DA	90.12	0.906	0.905	0.983
QUEST	89.25	0.899	0.889	0.982
NB	76.08	0.945	0.942	0.991

B - Segment dataset

Model	Accuracy	Recall	Precision	Specificity
NN	83.33	0.841	0.843	0.943
LR	80.85	0.819	0.815	0.935
SVM	76.6	0.777	0.768	0.921
DA	75.89	0.771	0.752	0.919
C5.0	75.53	0.766	0.755	0.918
CART	70.92	0.715	0.749	0.901
QUEST	67.02	0.679	0.694	0.888
CHAID	61.35	0.622	0.628	0.870
NB	12.41	0.655	0.660	0.907

C - Vehicle dataset

Model	Accuracy	Recall	Precision	Specificity
C5.0	67.67	0.368	0.490	0.912
SVM	60.4	0.369	0.518	0.891
LR	59.9	0.346	0.513	0.889
NN	57.39	0.337	0.335	0.888
QUEST	57.39	0.240	0.305	0.879
NB	56.14	0.354	0.409	0.885
CART	55.14	0.272	0.269	0.878
CHAID	53.38	0.304	0.305	0.874
DA	38.85	0.468	0.288	0.871

D - Red Wine dataset)

Model	Accuracy	Recall	Precision	Specificity
NB	64.29	0.673	0.635	0.613
LR	64.29	0.619	0.650	0.667
SVM	63.99	0.625	0.644	0.655
C5.0	63.99	0.607	0.642	0.661
NN	63.99	0.589	0.656	0.690
QUEST	63.1	0.619	0.634	0.643
CART	61.31	0.607	0.614	0.619
DA	60.71	0.607	0.607	0.607
CHAID	59.82	0.577	0.602	0.619

E -NHANES Dataset

Model	Accuracy	Recall	Precision	Specificity
C5.0	90.97	0.909	0.911	0.954
NN	75.03	0.747	0.763	0.872
CART	73.38	0.726	0.751	0.863
CHAID	73.18	0.731	0.738	0.863
LR	71.11	0.713	0.718	0.852
SVM	69.92	0.696	0.710	0.846
DA	68.29	0.684	0.711	0.838
QUEST	66.96	0.667	0.677	0.831
NB	64.05	0.642	0.636	0.817

F- House dataset

V. CONCLUSION

In this paper, classification algorithm have been implemented on nine datasets. The goal of the research was to evaluate the performance of the classification algorithms on both multiple and binary classification problems using a variety of performance metrics: classification accuracy, precision, recall, and specificity, lift charts gain charts.According to the experimental results, the C5.0 model proved to have the best performance. It performed better in many of the datasets used. Neural networks, naïve Bayes and logistic regression also performed well. However, there is no universally best

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 10, October 2013

learning algorithm. From the analysis none of the algorithms outperformed the others in every problem. The performance of classification algorithm depends on the performance matrix and the characteristics dataset.

REFERENCES

[1]Atlas, L., Connor, Park, J. , El-Sharkawi, D. , Marks , M. , Lippman, R., Muthasamy, A.Y. , " A Performance Comparison of Trained Multi-layer Perceptions and Trained Classification Trees". Systems, man, and cybernetics: proceedings of the IEEE international conference, 915-920 , 1991.

[2] Berardi, V. L., Patuwo, B. E., and Hu, M. Y. ,"A principled Approach for Building and Evaluating Neural Network Classification Models". DecisionSupport Systems, 233-246, 2004.

[3] Bhattacharyya, S., andPendharkar, P. C. ,"Inductive, Evolutionary and Neural Computing Techniques for Discrimination: A Comparative Study" ,1998.

[4]Breiman, L. , Friedman, J. H., Olshen, R. A., and Stone, C. J.,"Classification and Regression Trees". Wadsworth,Belmont,1984.

[5] Brown, D., Corruble, V., and Pittard, L. , "A Comparison of Decision Tree Classifiers with Backpropagation Neural Networks for Multimodal Classification Problems". Pattern Recognition, 26, 953-961.

[6]Burges, C. , "A Tutorial on Support Vector Machines for PatternRecognition". Data Mining and Knowledge Discovery. Kluwer AcademicPublishers. Boston, 1998.

[7]Caruana, R., andNiculescu-Mizil, A. , "An Empirical ComparisonofSupervised Learning Algorithms." Proceedings of the 23rd InternationalConference on Machine Learning,2006.