International Journal of Engineering &Technology, 7 (3.12) (2018) 411-414

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Logistic Regression and Data Analysis on Privacy Methods on Data Streams

P Chandrakanth[1]*, Anbarasi M. S[2]

[1,2]Pondicherry Engineering College, Department of CSE, Puducherry, India
**Corresponding Author Email: [1]chandrakanth@pec.edu, [2]anbarasims@pec.edu*

Abstract

The problem data privacy in streams is completely put in a myopic view by hitherto researchers. Research and experimentations have been well fortified on static data, in which predominantly spelled easy with approaches based on perturbation using random data values. Approaches based on large data sets and high dimension data sets are not adequate consequences. By using the phenomenon of autocorrelation of multivariable streams and their leveraging structures, identifying the suitable areas to add noise maximally preserves privacy and in a irreversible manner. Drift checking and ensemble classifier building is the basic requirements for privacy preserving data stream, which makes clear in experimentation with the support of sensitivity analysis. In this paper we present the results of experimentation at all the stages.

Keywords: *concept drift, Logistic Regression, data utility, data streams, data Privacy, Privacy Preserving in Data Mining (PPDM).*

1. Introduction

The design of experiments is the act of designing of information variation under conditions, which is hypothesized to reflect variations. The design introduces conditions that directly affect variations called quasi-experiments. Natural conditions that usual influence the variations are selected for observations.
The objective of an experiment in its simple form is prediction. A change is introduced based on preconditions for prediction. These are represented by independent variables or "input variables" or "predictor variables". The change is hypothesized to result the change in one or more dependent variables or "output variables" or "response variables". There control variable that are also identified in the experiment which are constant to prevent external factors affecting the results.
The selection of certain variables like, independent, dependent and control variables is the core tasks of experimental design. The planning of the delivery requirements of the experiment are under statistically optimal conditions given the constraints of available resources.
There are varied set of approaches for determining the set of design points (rare and particular combinations of the setup schemes of the independent variables) to be used in the experiment.

2. Logistic Regression

A renowned statistician David Cox developed Logistic regression in 1958.[2][3] The logistic model has a most commonly used variant called the binary logistic model which is implemented to estimate the probability of a binary response. It uses one or more predictor (or independent) variables (features) whereupon it acts like basis.
The presence of a risk factor increase is presented among the odds of a given outcome with respect to a specific factor. The probability of output in terms of input is the specialty of this model. It can be used as classifier, even though it does not perform statistical classification. In this model the challenge in experimentation is by choosing a threshold value which is a cutoff and classifying the inputs with probability greater than the threshold as one class, below as the other.

3. Logistic Regression for Data Privacy

Privacy-preserving with machine learning is an emerging research problem, due in part to the increased reliance on the internet for day-to-day tasks such as banking, shopping, and social networking. Privacy is a mandate for the fields of medical, financial and insurance being data digitized, stored and managed by many autonomous systems. The literatures on cryptography and security related to information processing pronounce the definitions of data privacy. The design of machine learning algorithms adheres to these frameworks to be explored to zenith. Notions of privacy are also introduced in many data-mining algorithms that are least formally justified.

4. Data Privacy in Data Streams

The problem data privacy in streams, has received surprisingly limited attention. Research and experimentations have been well fortified on static data, in which predominantly spelled easy with approaches based on random perturbation of the data values. As a matter of fact, streams pose additional challenges. Firstly, the analysis of the data has to be conducted exponentially, with a challenge of using limited processing time and buffer space, where the batch approaches are found to be unsuitable.

Second, the streams embody the characteristics and evolve timely changes over time. Approaches based on large data sets and high dimension data sets are not adequate consequences. By using the phenomenon of autocorrelation of multivariable streams and their leveraging structures, identifying the suitable areas to add noise maximally preserves privacy and in a irreversible manner.

5. Privacy through Ensemble Classification

From the values in the samples considering standard deviation and mean, a dichotomous categorical variable is tested. A dichotomous variable is predictive measure developed in each stage of the experiment. The total experiment has 7 stages:

- Predicting the size of sample
- Understanding the drift in the samples extracted from the streams
- Identifying the attribute that needs to be added to the ensemble classifier
- Understanding the sensitivity of the classified data sets
- Estimating the noise level for the sensitive attributes
- Estimating the tuple that need perturbation
- Testing the perturbed data

Predicting the size of sample

The size of sample is an activity of choosing the number of observations to include in a statistical sample. For any such empirical studies sample size plays a very important role to make the inferences sensible. The proportion is estimated based on the quality and characteristics of data sets. Wishing that all the instance families in the splice gene data set participate at least to a majority, we have a maximum sample size of 800 and minimum sample size of 70, out of which the results are proved better as accurate for a 480 sample size. But his could be judged with empirical study on the samples taken on the several observations taken on a probability of 1.

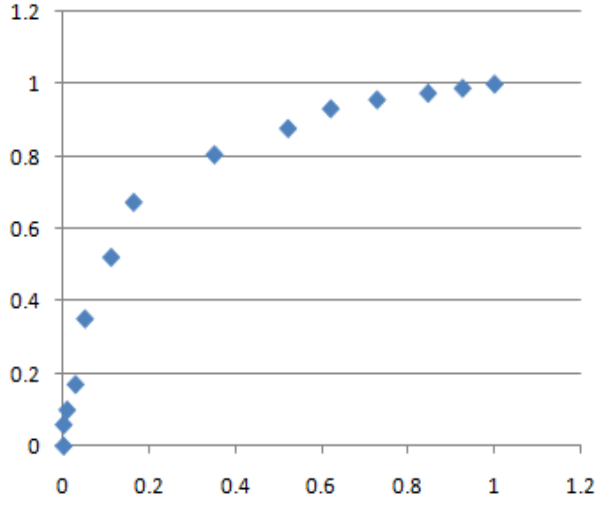

Fig. 1: Sample size prediction

Understanding the drift in the samples extracted from the streams

Though samples sizes are fixed the generator for splice gene datasets could generate only approximate sizes. The drift is identified in the samples in different classes of sizes in samples. The lowest size has all observations; the other sizes have only some significant observations. From the prediction probabilities the success and failure of identify the drift are noted and the quality of the algorithm is assessed by the curve.

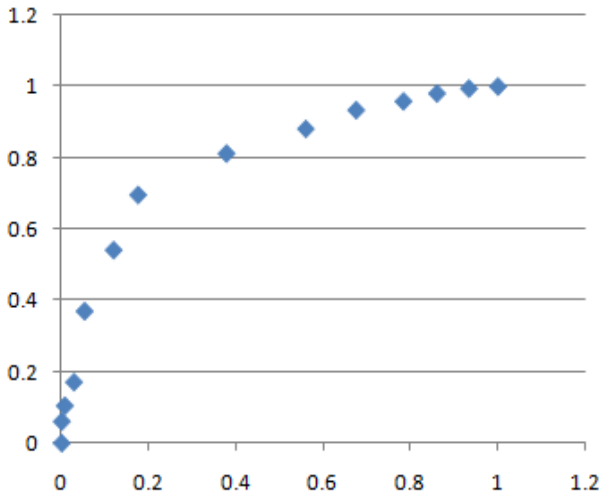

Fig. 2: Understanding the drift in samples

Identifying the attribute that needs to be added to the ensemble classifier

All attributes of all samples need not be significant enough to build criteria for the classifier; some are selected based on their frequency of access and their presence in the population. Selected attributes form the criteria based on the frequency and their content of instance sequences. Instead, the instance name is trimmed to an instance family and the instance family is the attribute building the criteria into the ensemble classifier. The poll for an instance family in the sample is observed from the experiments and their predictive probability and the probability of observations is mapped and illustrated in the curve.

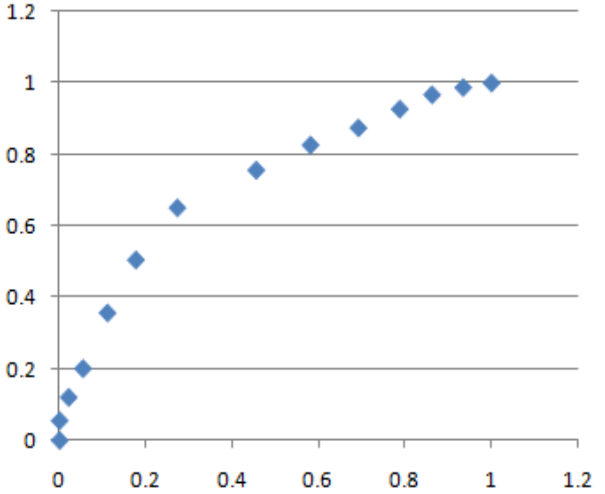

Fig. 3: Identifying attribute to become member in classifier

Understanding the perturbation according to sensitivity

In this phase of research work undertaken, the classified data sets are treated into perturbation. Allowing perturbation to all the tuples is a mere disturbance caused to the data sets, disallowing the analysis to complete extensively correct. The tuples which need the perturbation are identified by assessing the sensitivity and the tuples which are sensitive only are selected for perturbation. From the samples generated for each experiment and observations on the samples, the ensemble classifier is built and filtered records are tested for their sensitivity using the Generalized Proportionate Value developed in our work. The GPV is the sensitivity level required for the perturbation and becomes the seed value for the generation of the smooth noise. By adding the smooth noise to the selective tuples in the data sets makes

clear for the end analyst to estimate the originality of the data sets. As all the tuples are not perturbed, the other tuples reflect their originality guising among the perturbed ones. The probability of observations and predictions are mapped to call the success and failures with true positive ratio and false positive ration in the curve.

How much noise is required for the tuples to get perturbed is a very important and rather a challenging task. As said, all the tuples need not be perturbed, if so the perturbed ones, what is the suitable noise level to do so. However, the GPV value determines the sensitivity level the noise level is determined in the Perlin noise generator to determine the requisite disturbance. The observations of samples and predictive probability of the noise levels are drawn and treated in Kronecker sum for perturbation. The following three curves show the understanding the sensitivity of the classified data sets, estimating the required noise levels for perturbing tuples having sensitive attribute, estimating the tuples that really need perturbation.

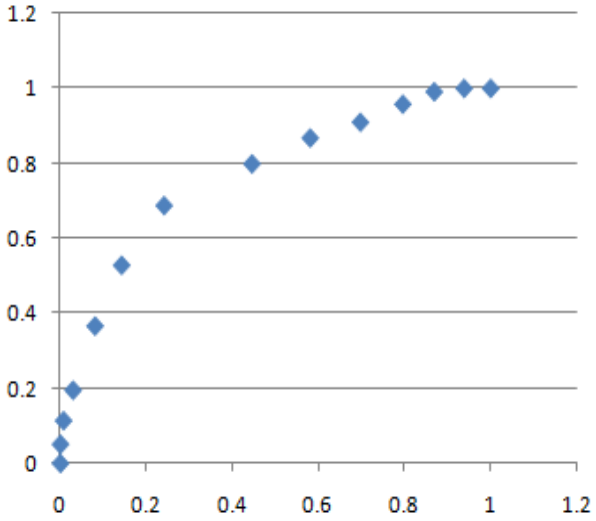

Fig. 4: Understanding the sensitivity of the classified data sets

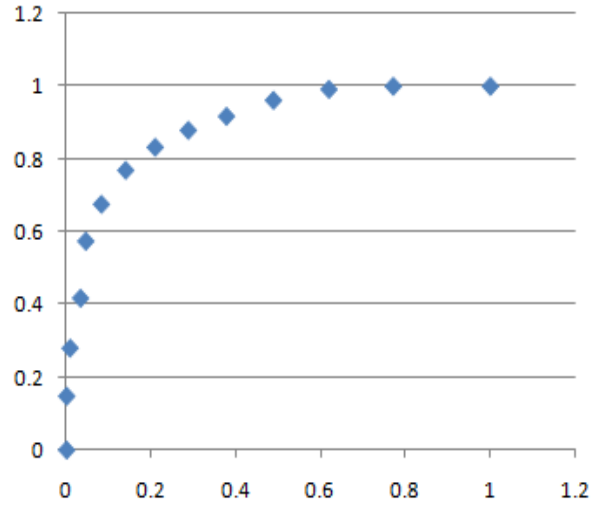

Fig. 5: Estimating the noise level for the sensitive attributes

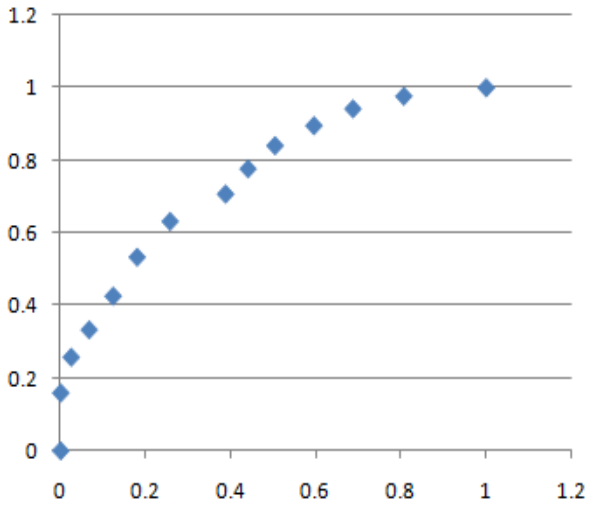

Fig. 6: Estimating the tuple that need perturbation

6. Testing the Perturbed Data

Testing the perturbed data needs an end user analyst environment, to check whether the results after perturbation and before perturbation are same. As far as concerned, the testing of results can be undertaken with data mining algorithms for clustering. The difference of clusters with outliers shall be tested. The clustering algorithms are tested on the WEKA tool. The WEKA has two important clustering mechanisms, namely DBSCAN and EM algorithms. The results of the two algorithms are found similar for both after perturbed data and before perturbed data.

Density-based spatial clustering of applications with noise - DBSCAN is an algorithm for data clustering proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996 This algorithm finds a number of clusters starting from the estimated density distribution of corresponding nodes. The two important parameters in this algorithm are distance and the min points. DBSCAN is the most common clustering algorithms cited in the scientific literatures.

Expectation Maximization - EM algorithm is also an important algorithm of data mining. This algorithm is used based on the results of k-means methods. This is an iterative method. This is used to find maximum likelihood or *maximum a posteriori* (MAP) estimates of parameters in statistical models. In this, the model depends on unobserved latent variables. The iteration alternates between performing an expectation (E) step, which computes exponentially the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximization of the expected log-likelihood found in the E step. The parameter-estimates are used to determine the distribution of the latent variables in the next E step.

One of the simplest unsupervised clustering algorithms is K-means (Macqueen, 1967). The procedure applies following a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. Defining k-centroids is the key area of computation for each cluster. The centroids are placed in a as pseudo centres, because different location causes different result.

Table 1: Comparative results of observations in DBSCAN, EM and K-Means algorithms for before and after perturbation of data

Samples	DBSCAN			EM			K-Means		
	BeforeP	AfterP	Result	BeforeP	AfterP	Result	BeforeP	AfterP	Result
70	2	2	1	3	3	1	2	2	1
120	3	3	1	2	2	1	3	3	1
320	4	3	0.75	4	3	0.75	4	4	1
400	5	4	0.8	5	4	0.8	2	2	1
440	3	3	1	4	3	0.75	3	3	1
480	4	4	1	3	3	1	4	4	1

485	2	2	1	2	2	1	3	3	1
540	5	4	0.8	3	3	1	2	2	1
600	3	3	1	4	3	0.75	3	3	1
640	2	2	1	5	4	0.8	2	2	1
660	3	3	1	2	2	1	3	3	1
720	5	4	0.8	3	3	1	2	2	1
800	4	3	0.75	5	3	0.6	4	4	1

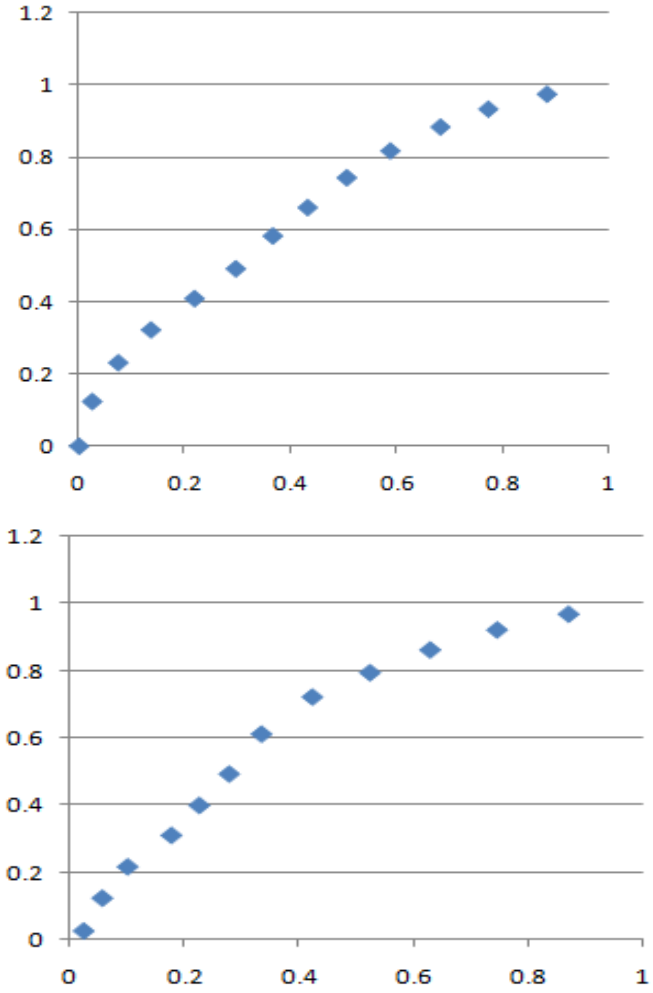

Fig. 7: Results of before and after perturbation in DBSCAN and EM clustering algorithms

7. Conclusion

The above ROC curves prove the consistency and correctness of the perturbation algorithm with clustering algorithms, that the perturbed data before and after are equally showing the productive results for DBSCAN and EM algorithms.For K-Means algorithms the results are exactly same as the above table indicates.

References

[1] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[2] Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54: 167-178. doi:10.2307/2333860.

[3] Jump up ^ Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20: 215-242. JSTOR 2983890.

[4] Charu C. Aggarwal and Philip S. Yu, "Privacy-Preserving Data Mining - Models and Algorithms", © 2008 Springer Science+Business Media, LLC. ISBN: 978-0-387-70991-8 [524 pages].

[5] Jaideep Vaidya, Chris Clifton and Michael Zhu, "Privacy Preserving Data Mining", © 2006 Springer Science+Business Media, Inc.

[6] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy Preserving Data Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, Pp. 1598, © September 2012.

[7] Aristides Gionis and Tamir Tassa, "k-Anonymization with Minimal Loss of Information", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No.2 pp.205, © February 2009.

[8] Murat Kantarcioglou and Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 16. No.9, pp.1025, © September 2004.

[9] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions on Knowledge Discovery and Data Engineering, Vol. 26. No. 4, pp.969. © April 2014.

[10] Xue, Yanbing, and Milos Hauskrecht. "Active learning of classification models with Likert-scale feedback." Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017.

[11] Guigo, Roderic, et al. "An assessment of gene prediction accuracy in large DNA sequences." Genome Research 10.10 (2000): 1631-1642.

[12] Boone, Harry N., and Deborah A. Boone. "Analyzing Likert data." Journal of extension 50.2 (2012): 1-5.

[13] Wu, Huiping, and Shing-On Leung. "Can Likert Scales be Treated as Interval Scales?—A Simulation Study." Journal of Social Service Research 43.4 (2017): 527-532.

[14] Cao, Xi Hang, Ivan Stojkovic, and Zoran Obradovic. "A robust data scaling algorithm to improve classification accuracies in biomedical data." BMC bioinformatics 17.1 (2016): 359.

[15] Bornholt, James, et al. "A DNA-based archival storage system." ACM SIGOPS Operating Systems Review 50.2 (2016): 637-649.

[16] Chormunge, Smita, and Sudarson Jena. "Efficient Feature Subset Selection Algorithm for High Dimensional Data." International Journal of Electrical and Computer Engineering6.4 (2016): 1880.

[17] Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." Advances in bioinformatics 2015 (2015).

[18] Tijmstra, Jesper, Maria Bolsinova, and Minjeong Jeon. "General mixture item response models with different item response structures: Exposition with an application to Likert scales." Behavior research methods (2018): 1-20.

[19] Hochbaum, Dorit S., and Philipp Baumann. "Sparse computation for large-scale data mining." IEEE Transactions on Big Data 2.2 (2016): 151-174.

[20] Göb, Rainer, Christopher McCollin, and Maria Fernanda Ramalhoto. "Ordinal methodology in the analysis of Likert scales." Quality & Quantity 41.5 (2007): 601-626.

[21] Koufakou, Anna, Justin Gosselin, and Dahai Guo. "Using data mining to extract knowledge from student evaluation comments in undergraduate courses." Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE, 2016.

[22] Michalopoulou, Catherine, and Maria Symeonaki. "Improving Likert Scale Raw Scores Interpretability with K-means Clustering." Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 135.1 (2017): 101-109.

[23] Jain, Y. Kumar, and Santosh Kumar Bhandare. "Min max normalization based data perturbation method for privacy protection." International Journal of Computer & Communication Technology 2.8 (2011): 45-50.

[24] Fernandes, Maria, et al. "Sensitivity Levels: Optimizing the Performance of Privacy Preserving DNA Alignment." bioRxiv (2018): 292227.

[25] Prasser, Fabian, et al. "Lightning: Utility-Driven Anonymization of High-Dimensional Data." Transactions on Data Privacy 9.2 (2016): 161-185.