DEVELOPING A MACHINE LEARNING REGRESSION MODEL RELYING ON INSTANCE-BASED LEARNING STREAMS FOR EFFICACIOUS DATA MINING

Pranshul Pahwa

Received: 17th November, 2019; Accepted: 29th December, 2019; Published: 07th January, 2020

ABSTRACT

Information mining is concerned about the investigation of information for discovering examples and regularities in the informational collections. Mathematical science is concerned about the assortment, investigation, understanding or clarification, and introduction of information. Measurements assume a significant job during the time spent information mining examination and similarly representation of information assumes a significant job in the basic leadership process. Occurrence Based Learning Streams is an occasion-based learning calculation used to perform a relapse examination on information streams. The calculation can deal with enormous information streams with less memory and computational force. The paper focuses on the execution of Instance-Based Learning Streams as an expansion to the enormous online examination structure for information stream mining to build up a relapse model. The investigation unhide the relapse examination could be performed on little informational indexes as well as on information streams as in the present case yet the technique for the examination will be diverse in the two cases. On account of the little informational index, the relapse models are direct, numerous and polynomial, while on account of information streams the whole examination is performed under the huge online investigation system by taking the two assessment parameters fundamental relapse execution evaluator and windows relapse execution evaluator. This discovering is first of its sort in writing.

1. INTRODUCTION

Advancement of technology has resulted in large storage of data. These large masses of data consist of some hidden information of strategic importance, which can be used for future analysis with effective decision making. The two important types of data analysis methods are classification and prediction. In the former case a model is constructed (classifier) to predict the categorical labels, in the latter case a model is constructed (predictor) to predict the continuous variables. Regression examination is a factual strategy which is generally utilized for expectation and determining. It is additionally considered as the piece of the AI process. It assumes a significant job in the forecast. The authors1 have clarified altogether about relapse investigation. It tends to be utilized to demonstrate the connection between at least one autonomous factors and a needy variable. Different techniques are developed to carry out regression analysis, namely, Linear regression, multiple regression, polynomial regression and ordinary least squares regression. These methods can be parametric or non-parametric in nature. Many tools are also developed to carry out regression analysis viz. miniTab, Gnumeric, PASW. Regression analysis can be performed very efficiently using MS-Excel also. Rapid advancement of the technology resulted in the storage of digital data has also increased very rapidly. The continuous arrival of data is referred to as data stream.

Network monitoring data, sensor data, web clicks, usage of credit cards weather forecasting data are few examples of data streams. The data streams are massive in nature and they arrive at a very high speed. Data mining techniques are not suitable for mining data streams and the data streams must be processed under very strict constraints of space and time. Gaber et al.2, Gaber and Gama3 and Ikonomovska4 have explained many aspects of data streams, different characteristics and many other special features of data streams. Data streams can be mined using Massive Online Analysis (MOA) frame work. Basic data mining techniques, classification, clustering and association rule mining can be performed on data streams using MOA framework. The present work mainly aims at regression analysis using IBLS in MOA framework. The paper is organised as follows: Section 2 mainly discusses on the related work in the area of regression modelling and need and importance of the problem;

Section 3 about the IBL stream learning Section 4 discusses about methodology about regression analysis in MOA; Experiments and results are presented in Section 5; Finally, section 6 is about conclusions and future work.

2. RELATED WORK

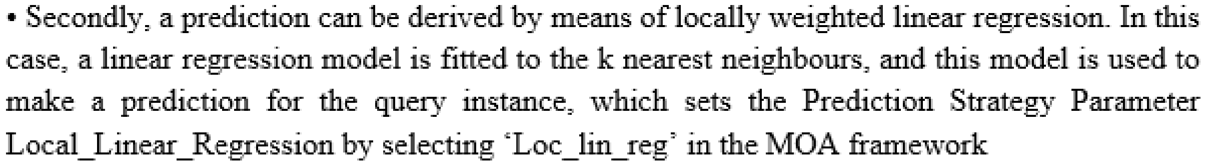
An exhaustive review of the writing uncovers that meagre writing is accessible relating to the present work. The work done in such a manner is informed in this segment. A large portion of the related work was found on the investigation and perception of information. The target of the examination and representation of information is to feature helpful data and bolster basic leadership. The authors emphasize the use of statistical approach in Edu data, generated from technical education system which consists of three stake holders namely Student, Faculty and Management. Authors have exhaustively studied linear5–7 regression analysis using PASW–18 statistical tool8 on Edu-Data and results were found to be very accurate and it is one of its kind on edu_data. Overall the study reveals that the regression model is developed on static data. As the technology is advanced the data collection is found to be massive in nature and such a data is referred to as data streams9 or massive data. Later study reveals that the possibility of regression modelling on massive data, as till now it was restricted only on a small sample of data. The author10 proposed a new method for mining large data sets using regression classes. It is defined as subset of the large data set which can be used for regression modelling. A large data set is treated as a mixture of many such regression classes. Fan11 consider block weighted least square estimators of the regression coefficients by minimizing the variances of the estimators and prove the asymptotic properties of the resulting estimators. They also indicate that the estimators make better interval estimation in terms of coverage probabilities than the usual least square estimators. The recent work carried out by Shakeer12, presents an approach to learning on data streams called Instance Based Learning Streams (IBLS). They have introduced the main methodological concepts underlying this approach by a mathematical model and implementation under MOA framework using a fuzzy model called as FLEXFIS13. MOA frame work was introduced by

Bifet et al.14–17. Subsequent citations Bifet et al.18 elaborates the use of IBL streams as extension in MOA framework. Therefore, the present investigation is aimed at developing a regression model using instance based approach using IBL streams on massive online analysis. The model uses two different prediction strategies Weighted Mean (W_MEAN_REG) and Local_Linear_Regression (LOC_LIN_REG) where one Adaptation_Strategy (adapt_k) and two parameter evaluation methods viz., Basic Regression Parameter Evaluator (BRPE) and Window Regression Parameter Evaluation (WRPE) and lastly the hold out evaluation method in MOA frame work are used. Thus the present work provides an excellent platform for future research.

3. REGRESSION ANALYSIS IN MASSIVE ONLINE ANALYSIS FRAME WORK

The present section emphasizes on the step wise explanation pertaining to the basic configuration method in MOA Framework. Shakeer12 explained the different possibilities of selecting the prediction and adaption strategies, different weighting methods and selecting the instance width which are basically used as fixed options for the experimental set up in MOA. This section provides the detailed explanation of the same. It is fundamentally utilized as an expansion to the system. It is comprehended from the system of relapse examination that, the objective quality is numerical and misfortune is ordinarily estimated regarding the outright or squared contrast between the anticipated yield and the genuine yield. Essentially, the MOA setup steps are additionally founded on a similar methodology. The forecast issues can be understood in two different ways.

• Firstly, the target value can be estimated by the weighted mean of the target values of the k neighbour instances. This prediction is obtained by selecting the option 'W_Mean_Reg', in the MOA frame work, which sets the Prediction Strategy parameter to Weighted_Mean_Regression.



• Secondly, a prediction can be derived by means of locally weighted linear regression. In this case, a linear regression model is fitted to the k nearest neighbours, and this model is used to make a prediction for the query instance, which sets the Prediction Strategy Parameter Local_Linear_Regression by selecting 'Loc_lin_reg' in the MOA framework

3.1 Prediction Strategies

In case-based learning, a forecast for the question example is acquired by joining, in one way or the other, the yields of the neighbours of this occurrence in the preparation information. The sort of total relies upon the kind of issue to be understood. MOA offers four diverse forecast plans, to be specific the Weighted Mode for grouping, the Weighted Median for ordinal characterization, and the Weighted Mean and Local_Linear_Regression for relapse issues. The present work uses only the prediction schemes.

3.2 Adaptation Strategies

Adapting to the size of the neighbourhood effectively contributes for the performance of an IBLS. Two different adaption strategies used in the evaluation are explained

as follows. The present work uses the first strategy:

3.2.1 Adapting_k

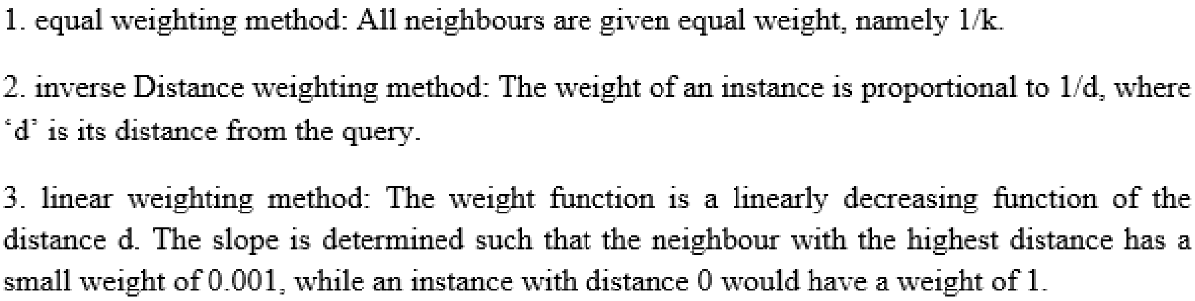
In this case, the size of the neighbourhood is controlled by the number of neighbours, k, and the IBLS algorithm adapts this value by continuously checking whether it appears beneficial to increase or decrease the current value by 1 for evaluation. This adaptation method is enabled in the GUI of MOA configuration model the by setting the parameter to adapt_K.

3.2.2 Adapting the Kernel Width

For this situation, the size of the area is controlled the weighting capacity or comparing part width. The calculation will at that point check whether variety in the estimation of bit width sigma by a specific rate gives off an impression of being useful for the assessment. This is activated by enabling the option adapt_sigma in the GUI of MOA configuration.

3.3 Weighting Methods

Five different Weighting methods are proposed.



1. equal weighting method: All neighbours are given equal weight, namely 1/k.

2. inverse Distance weighting method: The weight of an instance is proportional to 1/d, where 'd' is its distance from the query.

3. linear weighting method: The weight function is a linearly decreasing function of the distance d. The slope is determined such that the neighbour with the highest distance has a small weight of 0.001, while an instance with distance 0 would have a weight of 1.

4. Gaussian Kernel weighting method: Neighbours are weighted by centring a Gaussian kernel at the query instance.

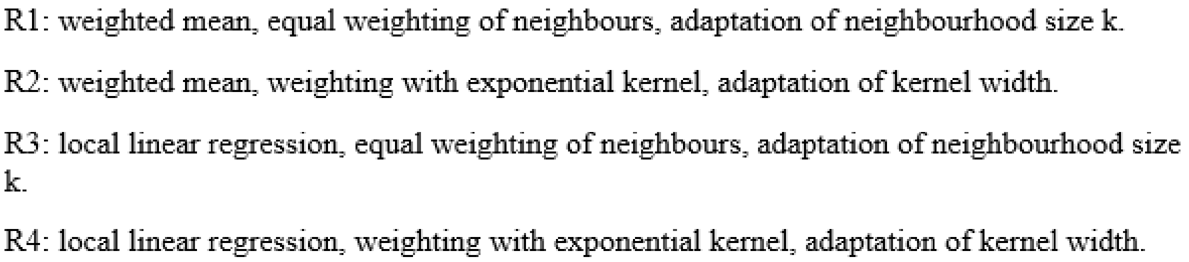
5. exponential Kernel weighting method: Neighbours are weighted by centring an exponential piece at the question example.

3.4 Selecting the Instance Width

Two important parameters used in the GUI of MOA for selecting the instance width are initial width and max_instance_base_size. Basically, the first parameter initial width defines the size of the initial set of instances on which IBLS produces the initial model. After checking the number of instances, the algorithm switches to its incremental mode of learning using a batch wise learning mode. Normally the default value for increment is 1000 instances. Second important parameter is max_instance_ base_size which upper-bounds the size of the training set which uses the default value as 5000 instances.

4. EXPERIMENTS AND RESULTS

Massive online analysis is the frame work used for mining data streams. The IBLS algorithm is implemented in MOA. IBLS is used as an extension to MOA. Configuration of the task launcher is configured for all the eight different data stream generators of MOA. (LED, HYPERPLANE, RANDOMRBF, WAVEFORM, AGRAWAL, SEA, STAGGER, RANDOMTREE) using both Basic Regression Performance Evaluator (BRPE) and Windows Regression Performance Evaluator(WRPE). The prediction strategies used are Local_Linear_Regression and Weighted Mean_ Regression. In the regression case, the IBLS are used in four different settings (while the rest of the parameters were again set to their default values).



R1: weighted mean, equal weighting of neighbours, adaptation of neighbourhood size k.

R2: weighted mean, weighting with exponential kernel, adaptation of kernel width.

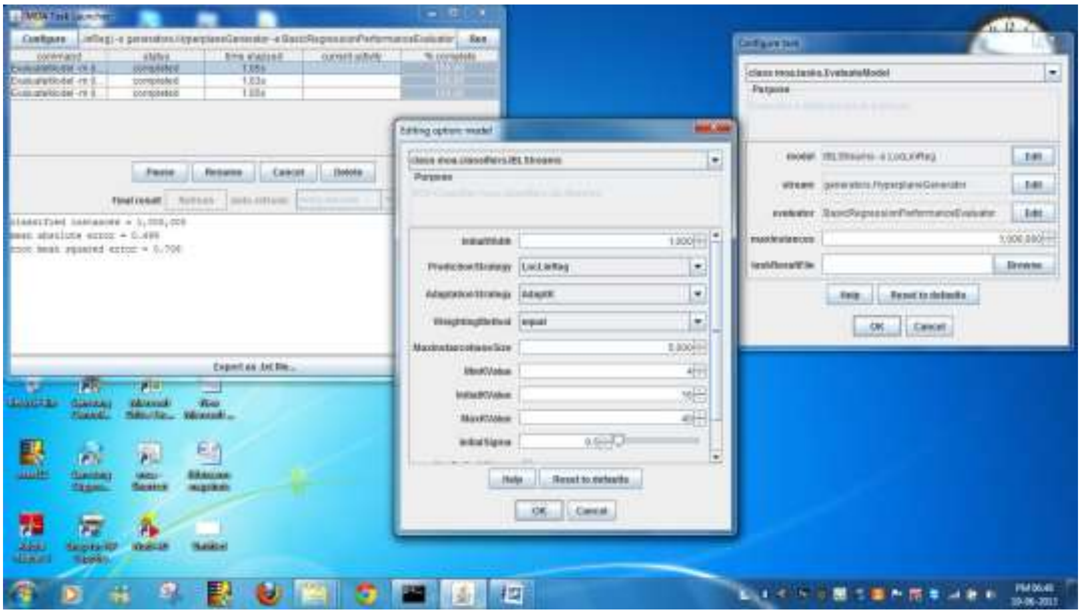
R3: local linear regression, equal weighting of neighbours, adaptation of neighbourhood size k.

R4: local linear regression, weighting with exponential kernel, adaptation of kernel width.

One sample setting for hyper plane data stream generator is shown in figure 1. From the above four settings it is R4. In addition to options of R1, BRPE is used. For the rest of the data stream generators the same type of settings is followed. The result window is shown in figure 2 and the results are tabulated in table1. The performance of IBLstreams learner on eight data stream generators in MOA framework is predicted in table 1. The experiment constitutes the evaluation of Mean Absolute Error (MAE) and Root mean square error (RMSE). The key features observed from the above table are presented below:

For any stream generator the prediction strategies w_mean_reg and loc_lin_reg and they happen to be same. For each prediction strategy the performance evaluators are Basic Regression Performance Evaluator (BRPE) and Windows Regression Performance Evaluator(WRPE) and they happen to be same. Eventually the values of MAE and RMSE are different for different data stream generators and minimum value is observed for stagger generator.

i.e. for the case of BRPE and prediction strategies w_mean_reg and loc_lin_reg, MAE=0.112, RMSE=0.334 and for the case of WRPE and prediction strategies w_mean_reg and loc_ lin_reg, MAE=0.103, RMSE=0.321. This proposes the presentation of IBLstreams is phenomenal on account of STAGGER generator A careful study of the writing (our papers) uncovers that the relapse investigation could be performed on little informational collections as well as on information streams as in the present case however the strategy for the examination will be diverse in the two cases. On account of a little informational index, the relapse models are straight, different and polynomial relapse, while on account of information streams the whole examination is performed under the Massive Online Analysis structure by taking the two assessment parameters fundamental relapse execution evaluator and windows relapse execution evaluator. This finding is first of its kind in literature and is quite interesting. Based on table 1, the results are presented graphically in Figures 3, 4, 5 and 6, for all the setting options R1, R2, R3 and R4 respectively and are self-explanatory.



command	status	time elapsed	current activity	% complete
EvaluateModel -m (I...	completed	1.05s		
EvaluateModel -m (I...	completed	1.03s		
EvaluateModel -m (I...	completed	1.03s		

```
classified instances = 1,000,000
mean absolute error = 0.499
root mean squared error = 0.706
```

class moa.classifiers.IBLStreams

Purpose

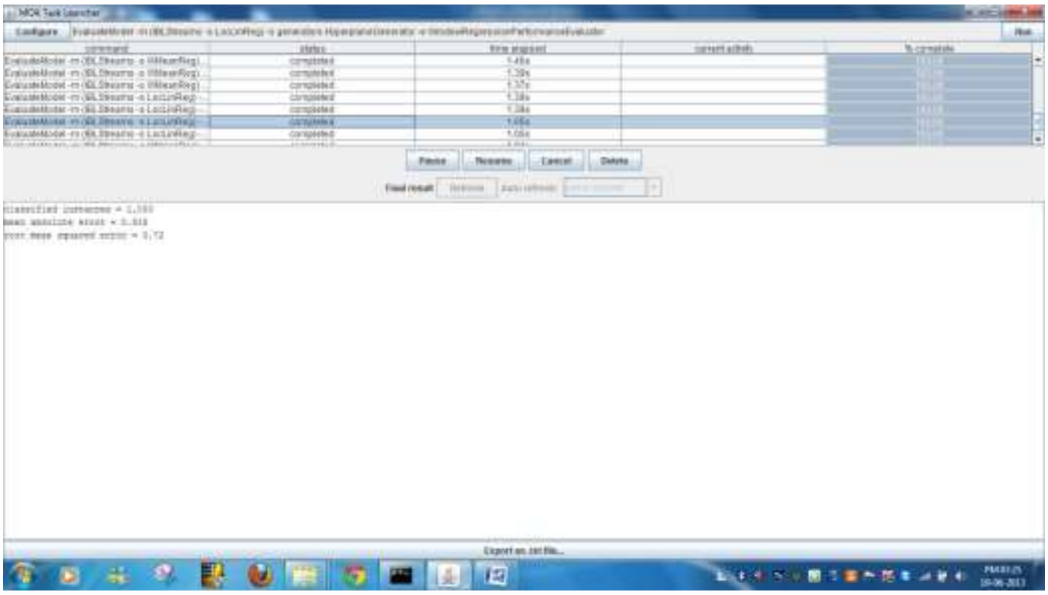
- InitialWidth: 1.000
- PredictionStrategy: LocLinReg
- AdaptationStrategy: AdaptK
- WeightingMethod: equal
- MaxInstancebaseSize: 5.000
- MinKValue: 4
- InitialKValue: 16
- MaxKValue: 40
- InitialSigma: 0.5

class moa.tasks.EvaluateModel

Purpose

- model: IBLStreams -s LocLinReg
- stream: generators.HyperplaneGenerator
- evaluator: BasicRegressionPerformanceEvaluator
- maxInstances: 1,000,000
- taskResultFile:

Figure 1. GUI for for Hyperplane Generator with R1 settings.



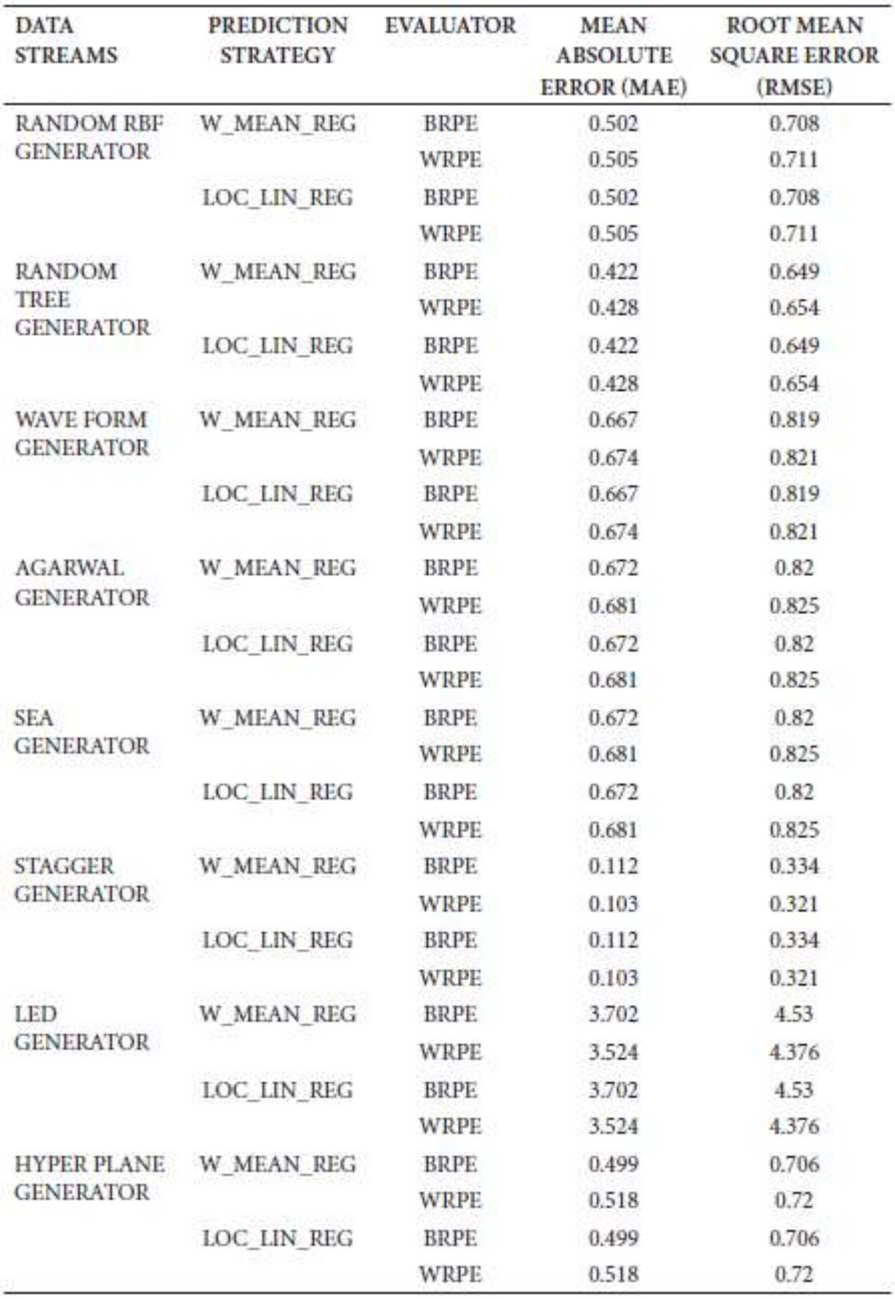
EvaluateModel -m (IBLStreams -s LocLinReg) -s generators.HyperplaneGenerator -e WindowRegressionPerformanceEvaluator

command	status	time elapsed	current activity	% complete
EvaluateModel -m (IBLStreams -s WMeanReg) ...	completed	1.45s		
EvaluateModel -m (IBLStreams -s WMeanReg) ...	completed	1.39s		
EvaluateModel -m (IBLStreams -s WMeanReg) ...	completed	1.37s		
EvaluateModel -m (IBLStreams -s LocLinReg) ...	completed	1.38s		
EvaluateModel -m (IBLStreams -s LocLinReg) ...	completed	1.38s		
EvaluateModel -m (IBLStreams -s LocLinReg) ...	completed	1.45s		
EvaluateModel -m (IBLStreams -s LocLinReg) ...	completed	1.06s		

```
classified instances = 1,000
mean absolute error = 0.518
root mean squared error = 0.72
```

Figure 2. Result window for Hyperplane Generator with R1 settings.

Table 1. Tabulation of Results for all data streams



DATA STREAMS	PREDICTION STRATEGY	EVALUATOR	MEAN ABSOLUTE ERROR (MAE)	ROOT MEAN SQUARE ERROR (RMSE)
RANDOM RBF GENERATOR	W_MEAN_REG	BRPE	0.502	0.708
		WRPE	0.505	0.711
	LOC_LIN_REG	BRPE	0.502	0.708
		WRPE	0.505	0.711
RANDOM TREE GENERATOR	W_MEAN_REG	BRPE	0.422	0.649
		WRPE	0.428	0.654
	LOC_LIN_REG	BRPE	0.422	0.649
		WRPE	0.428	0.654
WAVE FORM GENERATOR	W_MEAN_REG	BRPE	0.667	0.819
		WRPE	0.674	0.821
	LOC_LIN_REG	BRPE	0.667	0.819
		WRPE	0.674	0.821
AGARWAL GENERATOR	W_MEAN_REG	BRPE	0.672	0.82
		WRPE	0.681	0.825
	LOC_LIN_REG	BRPE	0.672	0.82
		WRPE	0.681	0.825
SEA GENERATOR	W_MEAN_REG	BRPE	0.672	0.82
		WRPE	0.681	0.825
	LOC_LIN_REG	BRPE	0.672	0.82
		WRPE	0.681	0.825
STAGGER GENERATOR	W_MEAN_REG	BRPE	0.112	0.334
		WRPE	0.103	0.321
	LOC_LIN_REG	BRPE	0.112	0.334
		WRPE	0.103	0.321
LED GENERATOR	W_MEAN_REG	BRPE	3.702	4.53
		WRPE	3.524	4.376
	LOC_LIN_REG	BRPE	3.702	4.53
		WRPE	3.524	4.376
HYPER PLANE GENERATOR	W_MEAN_REG	BRPE	0.499	0.706
		WRPE	0.518	0.72
	LOC_LIN_REG	BRPE	0.499	0.706
		WRPE	0.518	0.72

5. CONCLUSIONS

Regression analysis is a part of machine learning process, widely used for prediction, when numerical values are involved. Regression analysis assumes a significant job in demonstrating and investigating a few factors when the attention is on the connection between a reliant variable and at least one autonomous factors.

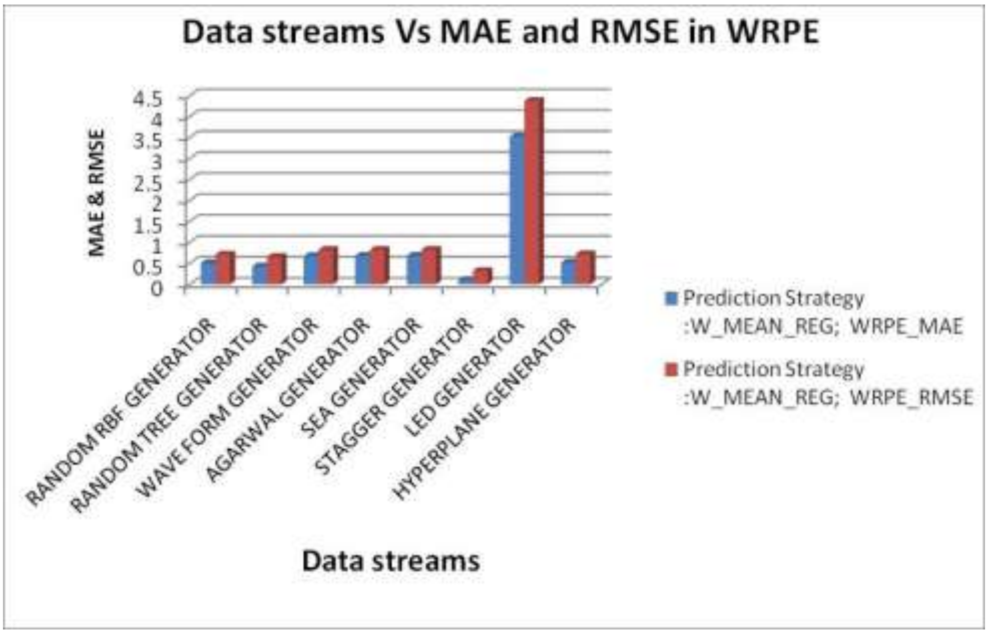


Figure 3. Graph of MAE and RMSE for Prediction Strategy = W_MEAN_REG, Performance Evaluator = BRPE.

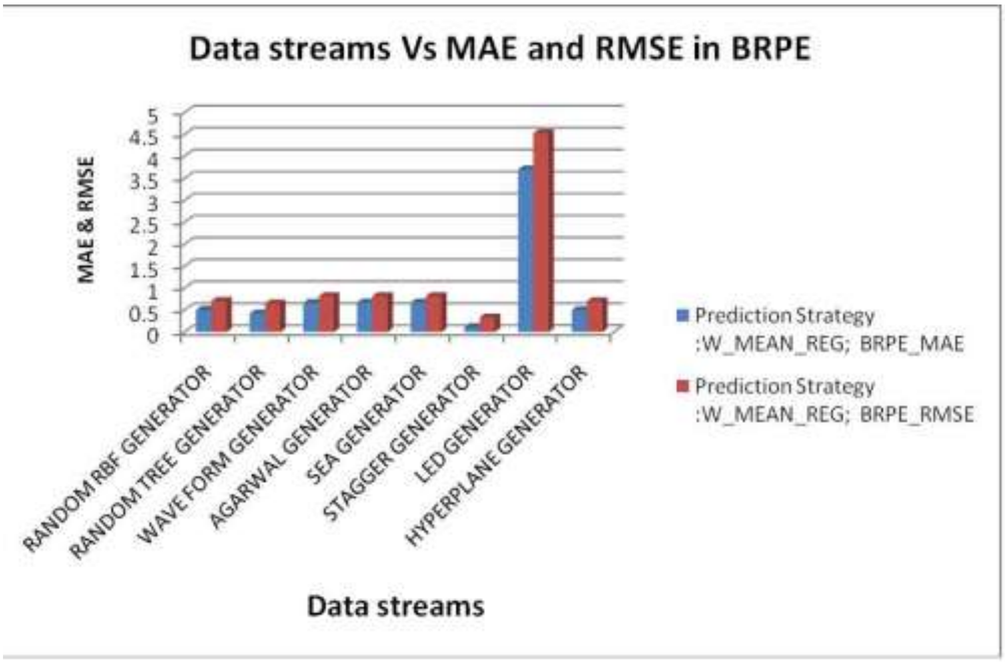


Figure 4. Graph of MAE and RMSE for Prediction Strategy = W_MEAN_REG, Performance Evaluator = WRPE.

Traditional DM techniques are not suitable for mining data streams because of their ubiquitous nature. Sophisticated techniques are required to mine data streams, Massive Online Analysis is one such frame work used for data stream mining. Basic data mining techniques, classification, clustering and association rule mining can be performed on data streams using MOA framework. The performance of IBLStreams learner on eight data stream generators in MOA framework is predicted. The experiment constitutes the evaluation of Mean Absolute Error (MAE) and Root mean square error (RMSE). The key features observed from the above investigation for any stream generator the prediction strategies are w_mean_reg and loc_lin_reg and evaluators are BRPE and WRPE. The performance of IBLStreams is excellent in the case of STAGGER generator.

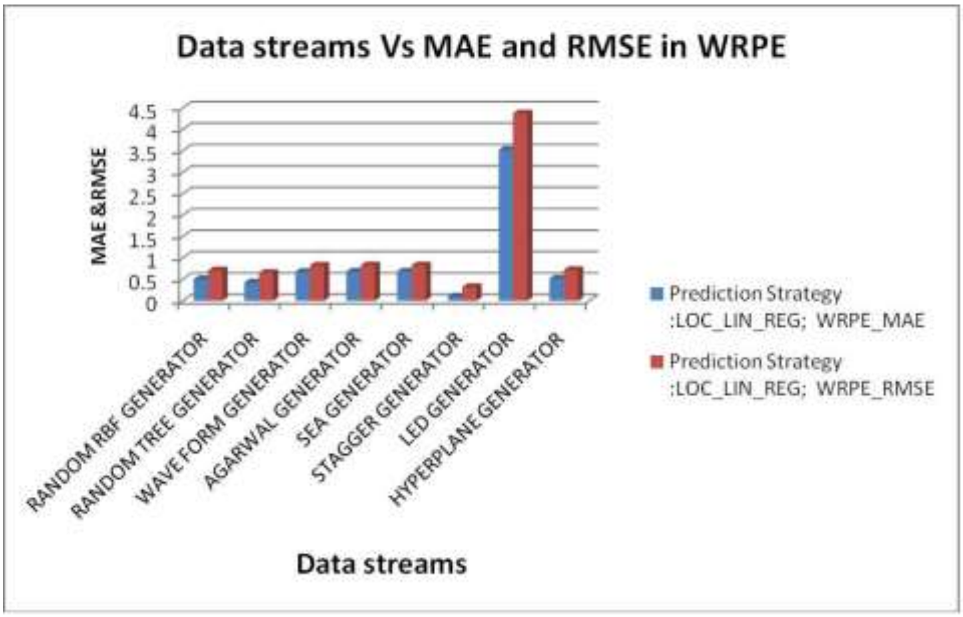


Figure 5. Graph of MAE and RMSE for Prediction Strategy =LOC_IN_REG, Performance Evaluator = WRPE.

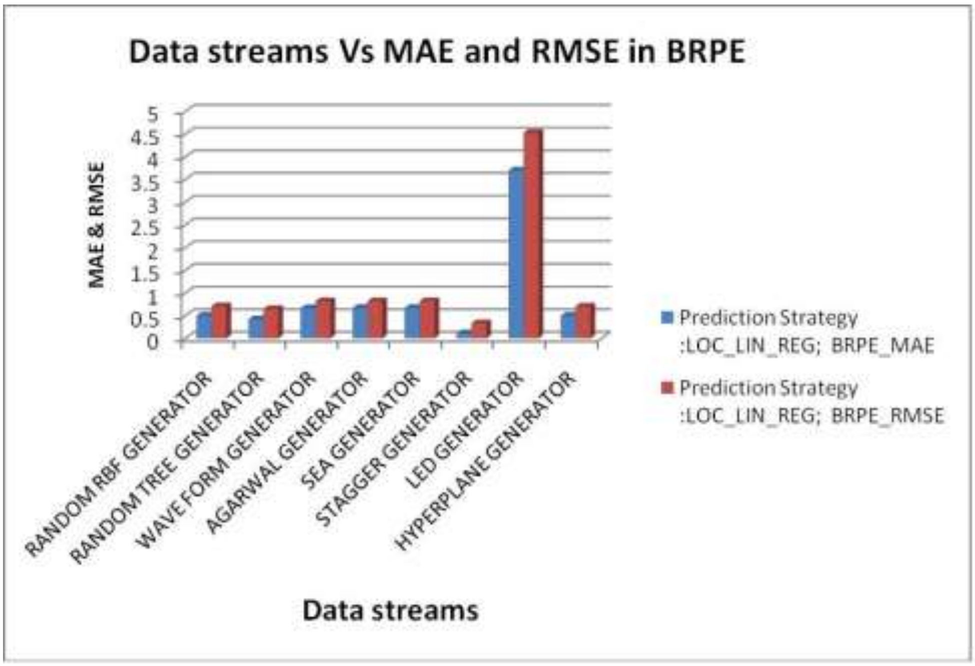


Figure 6. Graph of MAE and RMSE for Prediction Strategy =LOC_IN_REG, Performance Evaluator = BRPE.

At last, it is concluded that relapse investigation could be performed on small informational indexes as well as on information streams as in the present case yet the technique for the examination will be distinctive in the two cases. In the case of small data set the regression models are linear, multiple and polynomial regression, while in the case of data streams the entire analysis is performed under the Massive Online Analysis framework by taking the two evaluation parameters basic regression performance evaluator and windows regression performance evaluator. This finding is first of its kind in literature and is quite interesting.