Impact-Learning: A Robust Machine Learning Algorithm

Preprint · June 2020

CITATIONS
0

READS
415

3 authors:

Md Kowsher
Stevens Institute of Technology
64 PUBLICATIONS **184** CITATIONS
SEE PROFILE

Anik Tahabilder
Wayne State University
33 PUBLICATIONS **107** CITATIONS
SEE PROFILE

Saydul Akbar Murad
Universiti Malaysia Pahang
18 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Impact-Learning: A Robust Machine Learning Algorithm

Md. Kowsher
Noakhali Science and Technology University, Noakhali-3814, Bangladesh
+8801850884818
ga.kowsher@gmail.com

Anik Tahabilder
School of Engineering + Technology
Western Carolina University
Cullowhee, NC-28723, USA
+18282831045
tahabilderanik@gmail.com

Saydul Akbar Murad
Noakhali Science and Technology University, Noakhali-3814, Bangladesh
+8801521307785
saydulakbarmurad@gmail.com

ABSTRACT
The ultimate goal of this research paper is to introduce a robust machine learning algorithm called Impact-Learning, which is being used widely to achieve more advanced results on many machine-learning related challenges. Impact learning is a supervised machine learning algorithm for resolving classification and linear or polynomial regression knowledge from examples. It also contributes to analyzing systems for competitive data. This algorithm is unique for being capable of learning from a competition, which is the impact of independent features. In other words, it is trained by the impacts of the features from the intrinsic rate of natural increase (RNI). The input to the Impact Learning is a training set of numerical data. In this work, we used six datasets related to regressions and classifications as the experiment of the Impact Learning, and the comparison indicates that at outperforms other standard machine learning regressions and classifications algorithms such as Random forest tree, SVM, Naive Bayes, Logistic regression and so forth.

Keywords
Impact Learning, Classification, Regression, Machine Learning.

1. INTRODUCTION
Machine learning and data science procedures are being very significant in lots of fields. Smart spam detection, intelligence advertising systems, fraud detection, chatbot are excellent examples of machine learning applications. It is all about learning from the massive amount of data using some mathematical and statistical techniques.

The purpose of regression and classification is to develop a mathematical model in order to predict the dependent variable by analyzing a set of independent variables. In other words, we use a model to predict the value of Y when we know the value of X. Fundamentally, classification [1] is about predicting a label and regression[2] is about predicting a quantity and both are predictive modeling which maps a function from inputs to outputs called function approximation and can be categorized under the same umbrella of supervised machine learning[3]. In this learning, the algorithm is applied to develop the mapping function by mathematically analyzing the relationship between the input variables(x) to the output variable (y); mathematically can be considered as y = f(X). The objective of this type of problem is to calculate the mapping function (f) as accurately as possible such that whenever there is a new input data (x), the output variable (y) for the dataset can be predicted. To create such mapping function, there is a rush of strategies such as for linear, non-linear, and polynomial regression or in classification, SVM, KNN, Naïve byes, random forest tree, etc. Every strategy maintains a different methodology. Similarly, "impact learning" is a parametric statistical learning system and follows a different learning methodology. In demography, the rate of natural increase (RNI) [4] is a statistical concept and used in environmental science more. As data comes from reality, in order to make it autonomous, the RNI also can be used for building ML or data science applications. Here, the RNI is represented by the following expression:

$$\frac{dP}{dt} \approx rP \tag{1}$$

Here, we are considering r as RNI

Usually, the RNI's trend gets hindered for every element by a limitation, and we address that as the term carrying capacity(K). This concept can be better realized from the logistic growth model [5] of environmental science and statistics. Mathematically,

$$\frac{dP}{dt} = rP(1 - \frac{P}{K}) \tag{2}$$

However, every feature of a dataset follows the trend of RNI; on the other hand, there are more back forces by others on which the feature needs to be dependent. So, the target variable gets influenced by other features of the back forces, and we name that "Back Impact on Target (BIT)". Since the target feature relies on BITs, that is why every BIT also depends on the target feature.

Basically, the machine learning or statistical learning datasets derive from real sectors of target territories; consequently, they maintain the trend of the RNI. So, it will be a good way to generate the algorithm (Impact Learning) from the flow of RNI. Furthermore, this method learns from the effect of BITs, and in real life, every business sector has good competitors; the impact learning can be used in order to depict the competition among the competitors.

Moreover, to reveal the performance of impact learning, we used six types of datasets. We also showed the graphical and statistical comparisons among all existing machine learning algorithms such as Random forest tree, SVM, Naive Bayes, Logistic regression and so forth. The contributions are summarized as follows:

- Introducing a supervised machine learning algorithm named as impact learning.
- This algorithm follows the trend of RNI and learns from back impact related to other features.
- As it learns from the competition, it can be used to analyze the real competition.
- The experiments on the real dataset demonstrate the impact learning and compared with other remained methods.

In section II, we have mathematically introduced the impact learning algorithm. In section III, we have reported some examples

with their corresponding result and comparison and in section IV, we have concluded by discussing the result and mentioning some further work.

2. Mathematical Introduction of Impact-Learning

Basically, the logistic growth happens, whereas the rate of per capita growth reduces as population size reaches a maximum imposed by the limited resources, the carrying capacity (K). It can be written as:

$$\frac{dy}{dt} = ry(1 - \frac{y}{K}) \quad (3)$$

Again, the target features cannot walk towards the RNI's curve because of Back Impact on Target (BIT). If x is the back-impact variable, then x and y both keep their impact on each other's, and y keeps its impact on itself. So, we can rewrite the equation as

$$\frac{dy}{dt} = ry - w_y y^2 - wxy \quad (4)$$

Now, from the equation (1) and (2) we can write

$$ry\left(1 - \frac{y}{K}\right) = ry - w_y y^2 - wxy$$

$$\text{Or,} \quad r\left(1 - \frac{y}{K}\right) = r - w_y y - wx$$

$$\text{Or,} \quad y\left(\frac{r}{K} - w_y\right) = -wx$$

$$\text{Or,} \quad y = \frac{kwx}{r - w_y k}$$

If b is the bias, then we can get,

$$y = \frac{kwx}{r - w_y k} + b \quad (5)$$

If $x_1, x_2, x_3, \ldots\ldots\ldots, x_n$ are the impact features on y,

$$\text{Then,} \quad y = \frac{k\sum_{i=1}^{n} w_i x_i}{r - w_y k} + b \quad (6)$$

To determine the impact of x_k on target feature y (trained) and $y'$is the target feature, then we get from the equation

$$\text{Imp} = (y' - \left(\frac{k\sum_{i=1}^{n} w_i x_i}{r - w_y k} + b\right))^{\frac{2}{N}} \text{ if } i \neq k \quad (7)$$

Where, *N* is the size of the dataset.

If $X = [x_1, x_2, x_3, \ldots\ldots\ldots, x_n]$ and $W = [w_1, w_2, w_3, \ldots\ldots\ldots, w_n]$, then the equation (6) can be expressed as matrix dot product.

$$y = \frac{k(W^T . X)}{r - w_y k} + b \quad (8)$$

In order to illustrate the polynomial structure of the impact learning, the equation (6) can be expressed as:

$$\text{Then,} \quad y = \frac{k\sum_{i=1}^{n} w_i {x_i}^j}{r - w_y k} + b \quad (9)$$

Where $j > 0$

For the *N* size of data, the RNI and carrying capacity can be found from these defined functions.

$$r = \frac{\ln\frac{\max(y')}{\min(y')}}{N-1} \text{ and } k \geq \max(y')$$

But it is an excellent way to calculate the RNI (r) from the optimization techniques like gradient descent.

Figure 1 describes the flow chat of the proposed algorithm which is mathematically expressed in equation (6) above.

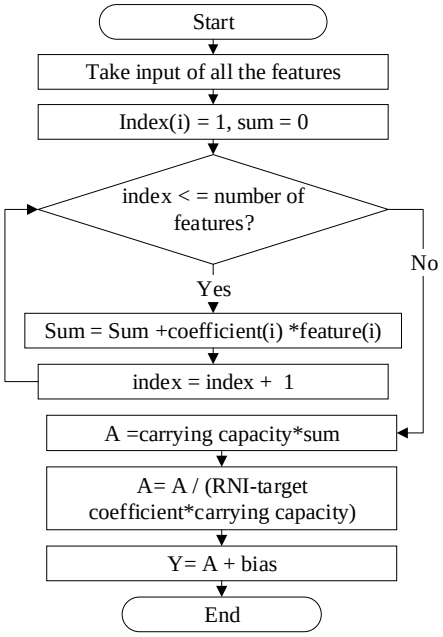

Fig.1 Flowchart of Impact-learning

3. Experiments

We describe a range of experiments to measurer the proposed "Impact Learning" algorithm. In this section, first, we introduce the datasets that we used in this research and describe the experimental setup. Then, we discuss the performance and result of our impact learning; and comparison with other machine learning as well as deep learning algorithms. We collected six different types of datasets, including the real competitive business, medical, Kaggle on regression and classification.

3.1 Experimental Setup

We have implemented both our propounded model and existing approaches such as regressions, classifications, and deep learning in Anaconda distribution in Python 3.7 programming language and executed them on a Windows 10 PC with an Intel Core i7 CPU (3.50GHz) and 32GB memory. Apart from sklearn, tensorflow, scipy, metplotlib, pandas are also used for reading data, preprocessing, splitting, training algorithms, developing neural network, graphical visualization, etc. In the following subsections, we recapitulate the experimental results that answer the above research questions. For the easy implication of code, we improved a python module of our model.

3.2 Training Model

The least-squares method is a widespread regression technique for analysis that can find the line for the best fit for a set of data, providing a visual demonstration of the relationship between the data points. Here the formula of the least-square method is:

$$b = \frac{\sum(x - \bar{x}) * (y - \bar{y})}{\sum(x - \bar{x})^2} \quad (9)$$

It is used in the training system of our proposed model.

3.3 Real Competition Analyzing and Prediction

Since the "impact learning" learns from the impact of others related features, it can be trained from the competitions of independent features. For this consequence, the real-life competition and prediction can be resolved through this method.

Data Collection. To analyze the competitive prediction, we collected the 10 years' mobile phone subscriber's data [6] from the Bangladesh Telecommunication Regulatory Commission (BTRC), where all the data of the Bangladeshi mobile phone companies were presented.

Performance. After the model training, we noted that all the coefficients of our fitted model. Let's our target agent is 'Robi' then the impact learning will

$$y = \frac{k}{r - w_y k}(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5) + b \quad (10)$$

Where x_1, x_2, x_3, x_4, and x_5 are respectively the features such as GP, Banglalink, Airtel, Citycell and Teletalk. If the algorithm is trained with the help of least square method or gradient descent, then the curve is fitted as figure. 2.

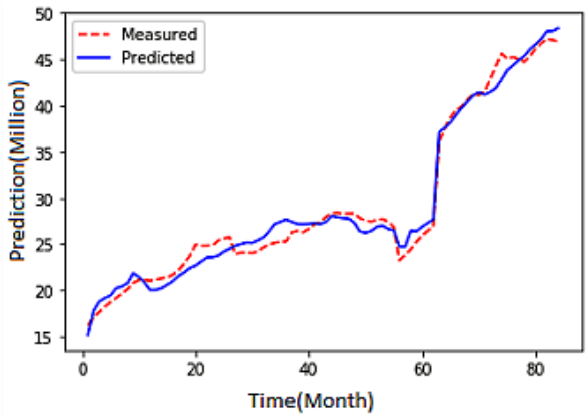

Fig.2. The fitting curve of Robi using Impact-Learning

To calculate the errors, we have taken the execution of mean square error. From the curve fitting, the following estimation of error is 1.20582 in a million units. Here all the coefficients of impacted features are given in table-1, and the measurements of impact are revealed in table-2.

Table 1. Robi's trained coefficients of impact learning

Coefficient's name	Value
Grameenphone	-0.171302
Banglalink	0.034011
Airtel	0.224212
Citycell	-0.940897
Teletalk	0.141534
Constent	0.000000
RNI	0.141534
Carrying Capacity	0.203432

Table 2. The impact of the features on Robi

Impact on Robi	value
Grameenphone	2100.22
Banglalink	25.7711
Airtel	60.7784
Citycell	25.8031
Teletalk	5.51507

Form table-2, we can realize that GP's impact (2100.22) is higher than the others. However, Teletalk (5.5150) placed the lowest position. This step aids in finding out the multicollinearity or redundancy and the biggest competitor of the target variable. As from the aspect of business competition analyses, we can prominently prove; GP is the biggest competitor of Robi, which is about 84 times Citycell or Banglalink and 35 times Airtel.

Table 3. Comparison of MSE between impact learning and multivariable regression of test set.

MSE of Proposed Mode	Impact Learning	Multivariable Regression	ANN
Grameenphone	1.0521	1.0713	1.0614
Robi	1.2058	1.2862	1.1989
Banglalink	1.3812	1.4810	1.3817
Airtel	1.6251	1.8720	1.7715
Citycell	1.9201	1.91	1.9309
Teletalk	2.3725	2.4024	2.4001

From the table-3, a better pictogram of errors can be depicted by the comparison between impact learning and usually machine learning & ANN in multivariable regression. The impact learning has revealed a stander result of predicting like the existing techniques. However, the existing techniques have a drawback for illustrating the competition and effect of features. At this point, impact learning makes cognition as learning from competitions.

3.4 Solving Classification problem using IL

Classification is the process of categorizing unknown data into a desired and distinct number of classes in which one can assign a label to each class.

After being successful for in competition analysis, we implemented our algorithm for solving classification problems. We have implemented the proposed algorithm to solve classification problem for different datasets including medical, Kaggle, Gender detection and so one. The first dataset was "Pima Indians Diabetes Database" [7] that was collected from the National Institute of Diabetes and Digestive and Kidney Diseases, India. Our proposed algorithm classified the dataset with reliable accuracy as shown in table 4.

Table 4. Result comparison for Diabetes detection

Name	Accuracy	Precision	Recall	F1	C.kappa	AUC
RF	0.79	0.69	0.57	0.62	0.48	0.82
SVM	0.79	0.70	0.55	0.61	0.47	0.72
NB	0.79	0.67	0.61	0.64	0.49	0.84
LR	0.82	0.76	0.61	0.68	0.56	0.86
K NN	0.79	0.68	0.63	0.65	0.51	0.77
DT	0.70	0.51	0.63	0.57	0.35	0.68
LDA	0.82	0.76	0.61	0.68	0.56	0.86
IL*	0.83	0.76	0.61	0.69	0.58	0.87

From the table 4, it is clear that the impact learning algorithm provides comparatively better than all the existing algorithms, with both the highest accuracy and F1 score.

For this problem, the threshold obtaining from impact learning is 0.43 as shown in the figure 3. Impact learning does not follow the aptitude of the probabilistic interval [0,1] for binary classification. If we transfer the rate the function (equation 9) through sigmoid activation, it can be obeyed probabilistic interval [0,1]. Apart from, it is a better idea to figure out the threshold bound with the searching technique for the lowest cost function.

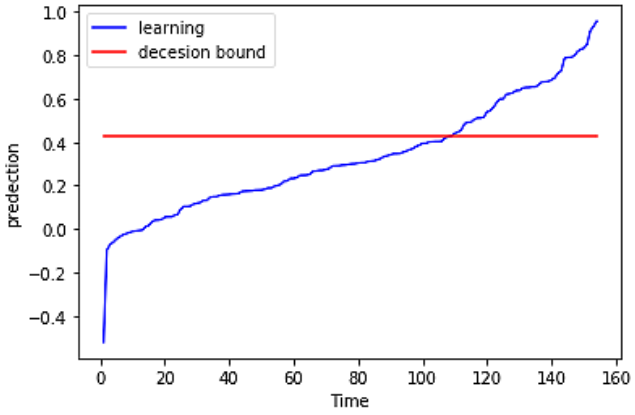

Fig.3. Threshold determination

We also applied the proposed algorithm to model a gender detection system [8], and the result is listed below in the table5.

Table 5. Gender detection

Name	Accuracy	F1	C.kappa	Recall	Precision	AUC
RF	0.86	0.87	0.66	0.85	0.87	0.90
SVM	0.87	0.88	0.71	0.86	0.87	0.91
NB	0.63	0.64	0.48	0.64	0.62	0.67
LR	0.87	0.86	0.69	0.86	0.88	0.90
KNN	0.85	0.86	0.71	0.84	0.86	0.89
DT	0.86	0.86	0.72	0.87	0.85	0.89
LDA	0.86	0.85	0.68	0.87	0.85	0.88
ANN	0.85	0.85	0.75	0.84	0.86	0.88
IL*	0.91	0.90	0.76	0.89	0.91	0.94

We also applied the proposed algorithm to select the best medicine for type-2 diabetics patients. The dataset was collected from the Noakhali Medical College, Bangladesh, consisting of 9483 samples and 14 symptoms per sample [9,10].

Table 6. Result comparison for type-2 medicine selection

Name	Accuracy	F1 Score
Logistic Regression	0.796524976	0.77923750
SVM	0.732950236	0.739735108
Naive Bayes	0.814529293	0.810929135
K-NN	0.815414013	0.846528487
LDA	0.795299207	0.780602597
Decision Tree	0.806335432	0.789429189
Random Forest	0.814352273	0.800348735
ANN	0.828152866	0.802268685
Impact Learning*	0.819143622	0.807832624

In addition, we applied the proposed algorithm to find the most appropriate mode of childbirth for a pregnant mother. This dataset was collected from the Tarail Upazilla Health complex, Tarail, Kishorganj, Bangladesh, and it contains medical information of 13,527 pregnant women in 21 different fields for each pregnant woman [11]. Table-7 describes the result comparison for various algorithm that was applied to predict the mode of childbirth.

Table7. Result comparison for childbirth mode prediction

Name	Accuracy	F1 Score
Logistic Regression	0.842675159	0.832870920
SVM	0.817197452	0.806994763
Naive Bayes	0.874522293	0.870958195
K-NN	0.855414013	0.846528487
LDA	0.835839207	0.820602597
Decision Tree	0.836305732	0.829431859
Random Forest	0.864352273	0.850348735
ANN	0.868152866	0.862268685
Impact Learning*	0.900189172	0.897871741

3.5 Multivariate linear Regression Problem

In machine learning, regression models attempt to predict a target in a continuous space. Curve fitting is robustly being used by the data scientist to describe experimental data in the domain of machine learning. Since forecasting single variable linear or multivariate linear regression depends on all ingredients of the dataset but redundancy, there is created a competition field of features impact. So, the impact-learning can be applied to regression or prediction value. Here, we implemented IL to the well-known concrete strength [12] dataset, which is a multivariate linear regression dataset from Kaggle includes data of civil engineering aiming to indicate the compressive strength.

The linear regression model is usually well-suited for the data; otherwise, a multivariate linear model is more appropriate. The following example shows a fairly multivariate linear regression; consequently, we used the sklearn multivariable linear regression has been exerted. Nevertheless, ‘impact learning’ has been trained with the least square optimization method. The coefficient we got after optimization is listed in table 8.

Table 8. Comparison of coefficients of the features

Name	Regression	IL*
Cement	12.470751	-5965.344
Blast Furnace Slag	9.601679	-4592.892
Fly Ash	5.876634	-2811.002
Water	-3.109428	1487.586
Superplasticizer	1.882144	-900.3159
Coarse Aggregate	1.729819	-827.3380
Fine Aggregate	1.823053	-871.9034
Age	7.117277	-3404.624
Bias	0.000014	36.10087
Carrying Capacity	None	162.96927
Intrinsic Rate	None	0.004330
Impact of target	None	478.3604

From the coefficient table 8, we can conclude that impact learning takes the opposite sing of multivariable as it fits curve from the back force. Here, the coefficient of impact learning has been displayed as the times of 100. However, the total mean square error of impact learning regression and multivariable regression are 95.105804015332 and 95.127565682425 as shown in figure 4 and figure 5, and it could be presumed a standard error of impact learning regression comparing to a multivariable regression.

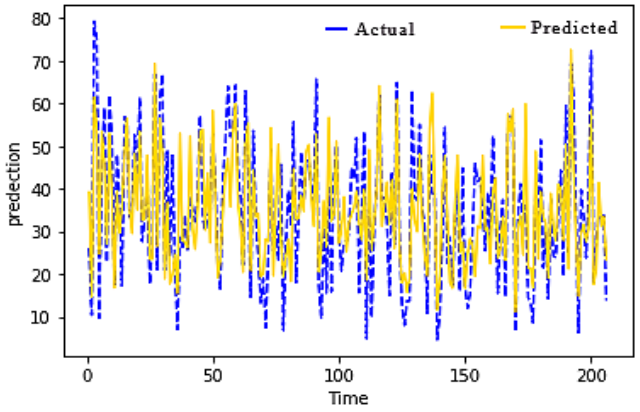

Fig.4 Result of sklearn's multivariable regression

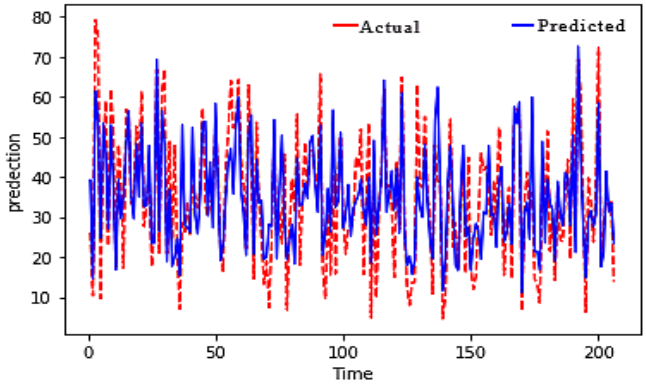

Fig.5 Result of impact learning's multivariable regression

4. Conclusion and Future Work

To put it into a nutshell, we have introduced a unique machine learning algorithm for resolving the regression and classifications problems. The principal technique of this method is the system of learning from RNI and the impact of other features like a competition. As it learns from the impact, it can be used for analyzing competition like the effect of other agents.

In the future, we plan to use the impact-learning the resolving of NLP problems, connecting to other machine learning and deep learning algorithms for better performance. Besides, this will be trained with the help of back-propagation & gradient descent instead of the least square method. We also have a scheme to apply this model for analyzing and forecasting value among the competitors in economics, business, learning machine, etc.

5. REFERENCES

[1] Deng, Zhenyun, et al. "Efficient kNN classification algorithm for big data." Neurocomputing 195 (2016): 143-148.

[2] Chen, Rong, et al. "Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm." Applied Soft Computing 26 (2015): 435-443.

[3] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.

[4] Taberner, Ana, et al. "Estimation of the intrinsic rate of natural increase and its error by both algebraic and resampling approaches." Bioinformatics 9.5 (1993): 535-540.

[5] Chen, Xianyao, et al. "The increasing rate of global mean sea-level rise during 1993–2014." Nature Climate Change 7.7 (2017): 492.

[6] http://www.btrc.gov.bd/telco/mobile [Accessed 10 Jan 2020]

[7] https://www.kaggle.com/uciml/pima-indians-diabetes-database [Accessed 1 Jan 2020]

[8] Kowsher, Md, et al. (In press)"Gender Identification from Bangla Name Using Machine Learning and Deep Learning Algorithms."Emerging Technologies in Data Mining and Information Security. Springer, 2020.

[9] Kowsher, Md, et al. "Type 2 Diabetics Treatment and Medication Detection with Machine Learning Classifier Algorithm." *Proceedings of International Joint Conference on Computational Intelligence*. Springer, Singapore, 2020.

[10] Kowsher, Md, et al. "Prognosis and Treatment Prediction of Type-2 Diabetes Using Deep Neural Network and Machine Learning Classifiers." *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019.

[11] Kowsher, Md, et al. (In press) "Machine Learning Based Recommendation Systems for the Mode of Childbirth" *2020 2nd International Conference on Cyber Security and Computer Science,2020.*

[12] https://www.kaggle.com/maajdl/yeh-concret-data [Accessed 02 Feb 2020]