ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 05, May 2025

# A Study on Dataset Drift Mitigation using Stream Based Ensemble Approach

[1] Mr. S. Thivaharan, [2] Sabarieswaran S, [3] Sudharshan K, [4] Lelith T

[1] Assistant Professor Department of CSE, KPR Institute of Engineering and Technology, Coimbatore, India
[2] [3] Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India
[4] Electronics and communication Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India
Email ID: [2] Sabarieswaransabari3@gmail.com, [3] sudharshanjr609@gmail.com, [4] lelith434@gmail.com

***Abstract*— Modern Knowledge systems are more towards taking immediate response working in real-time dynamic data streams. The infusion of dynamic ever changing stream of data is highly replacing the traditional static datasets, which in one way or the other, might have out-dated. By the time the static datasets complete their trail of training and testing and move for prediction, due to increased surge in dynamic infused data streams, the prediction model may not demonstrate the accuracy as expected. These dynamic streams are often challenged for their non-stationery environments. This research article performs a study over the stream datasets using the ensemble based model development. Many current researchers propose and argue that ensemble mechanism has the demerit of over-fitting and high reliance over a particular category of algorithm. The concept of drift driven strategy is proposed with the string backing of regression based metrics. Further this article also analyses the ensemble-variance tradeoff from the perspective of continuous stream of data.***

***Index Terms:*** *Dynamic data streams, over-fitting, ensemble-variance tradeoff, drift driven strategy, non-stationery environment.*

## I. INTRODUCTION

Ensemble learning uses multiple machine learning models for the same algorithm which try to make a better prediction. A specific dataset, instead of relying upon a particular algorithm, is fed as input to multiple machine learning models, and these models work over the dataset in their own procedural way and produces a combined prediction outcome. Figure – 1 below depicts the overall architecture of ensemble [1] based learning

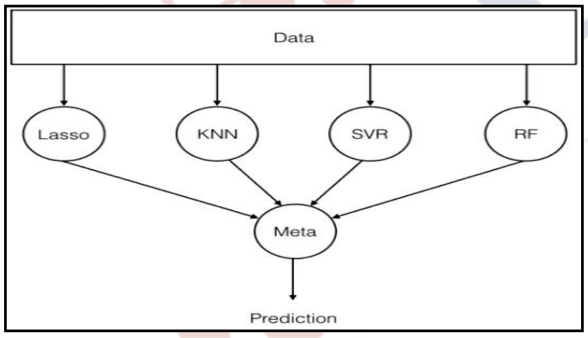



**Figure 1.** Overall architecture of Ensemble learning

Using a single algorithm for training the dataset might end up with over-fitting of the trained model. So the better practice is that, breaking down the overall model into three primitive components: Bias, variance and Irreducible error. As every model has merits along-side the weaknesses, this way of combining the effort of multiple algorithms helps in mitigating the error margin [2] in the final prediction classes. Out from the above mentioned primitive errors, the irreducible can't be handled all alone compromising the bias and variance.

The figure below better explains the bias and variance trade-off. The low bias and low variance scenario makes the final prediction more close to the actual expected outcome as well the model works as expected. The low bias and high variance scenario yields the prediction outcome to be situated around the expected and reality results but even then the predictions are spread around. The high bias and low variance makes the prediction as a cluster but even then the predictions are far from the actual expected reality class. The high bias [3] and high variance [4] not only makes the wrong prediction but the entire outcome is spread across un-uniformly in a different cluster. Figure-2 below represents the bias variance trade-off projecting the scattered nature of the data points.

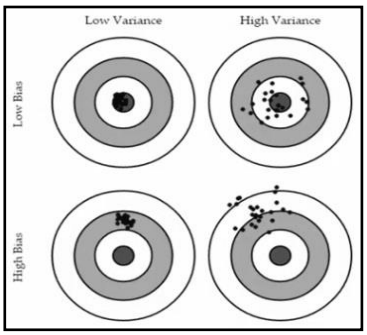



**Figure 2.** Bias – variance trade off comparison

Considering the below mentioned figure, after a certain point of time, if we start decreasing the bias, then the variance starts increasing. It is well understood that bias and variance are reciprocal to each other. In a nutshell, the addition of model complexity provokes the models variance and compromises the bias perimeter. The challenging task for any

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 05, May 2025

training model is to find the best spot where the bias-variance trade-off [5] is optimal. Figure-3 below depicts the variation of the model complexity as against the error probability.

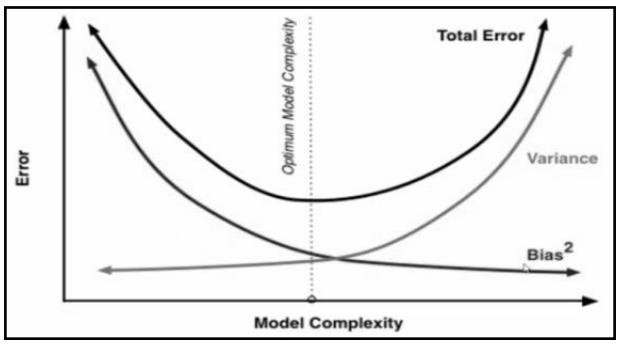



**Figure 3.** Model complexity vs. Error

The bias error is useful to quantify how much on an average the predicted values are different from the actual values. Low bias model are better performing models as compared to the high bias models which usually under performs. In contrast the high variance model yields the over-fitting scenario. After the optimal point where the bias and variance are cancelling each other, trying to increase the model complexity might make the model under consideration less dependable.

Ensemble techniques falls under one of the following category: Max voting, Averaging, Weighted average, stacking, blending, bagging and boosting. This research article concentrates on the bagging and boosting ensemble techniques. In bagging based ensemble category, bagging meta-estimator and Random forest algorithms play their quantification role. The following algorithms: Adaptive Boosting, Gradient Boosting [6], XG Boosting [7], Light GB [8] and CatBoosting [9] techniques are available.

In max voting, generalized classification optimality is used; as well the prediction is done for each of the available data points. Each of the datasets is deliberately attempted with the prediction cycle. The majority of the higher with standing accuracy yielding component is chosen for the further enhancement and scalability of the predictor.

In average voting technique, the dataset is trained with the help of multiple regression models, the outcome of the raining phase is mitigated by way of comparing and averaging the multiple verses of checkpoints produced at each of the dataset training intervals.

Ensemble predictions are always taken from the averaging of member prediction accuracy. A class label is calculated from the root class probability using the “argmax” and “summed probability tags” of the classifiers. Equal weighted contribution is made to final prediction attainment. The random chance of skill based model performance is known to be the “Worst –case mitigator”.

## II. LITERATURE SURVEY

**Hansen and salamon (2010) [10],** theoretically introduces the multi-dimensional spaces and correlates the earned probabilistic variations with the stochastic discrimination. Stochastic models are highly embedded neural layers targeting towards the classification tasks. Eventually these models prove to the promising solution providers in the random sub spaces with high degree of inter-connected separation points.

**Wolpert and Xu (2012) [11]**, introduced the generalized error data inhibition ratio by incorporating the concept of multi-layer Bayesian and dempster-shafer formalism classifiers. The above mentioned classifiers are out shining the available three types of classifiers: voting, argmax and generalized Bayesian classifiers. These models are capable of encaging with the normalized ensembles with elevated dataset points comprising of the feature derived ensemble metrics.

**Ho, yoav freund and Robert.E (2015) [12]**, claims that a small number of neural networks with sufficiently combined rationale to pin point the estimation correlation. This model is a fully connected cognitive based neural layer based derivative. A truly weighted representative training data set, where earliest and unlabeled ensembles are covered for the later analysis and mitigation of the error generation ratio. The numbers of filters that can be made with respect to extracted hyper features are aligned with kernel mapping parameters. The similar work based on a framework is been developed by **perrone and cooper (2016**) [13] in their ensemble generalized unbiased layers.

**Tumer and Ghosh (2016) [14],** talks about the practical emphasis on under-regularized kernels by way of depriving the over-fit dataset, allowing the variance-efficacy. This article also examines the performance equivalence of the majority used genetic voting techniques such as weighted voting, “Lamsian voting” and bi-directional Bayesian formalism. Individual classifiers with least navigational accuracy with feature space sampling with close correlated classifier rank mechanism with the worst and worst fitting dataset are measured for their voting complexity behavior.

**Kittler et al (2018) [15]** developed a combined framework with primitive fusion scenarios namely distinct representatives and fusion opinion representatives. The non-linear posteriori and the estimated resilient errors with the class prediction probability are summed up for the better plausibility. The benevolent sum-rule is high with scaled up product rule with the presence of both the distinct and shared variance representations. Random construction and effectiveness estimation for the voting scheme under investigation has the number feature and classification modeling and vectors simulations. The incorrect labels with the sequence of broad spectrum of data coarseness are trimmed with the auto selection of the RELU layers in their proposed work.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 05, May 2025

**Jain et al (2020) [16],** proposes a classifiers combination consisting of median, product and nearest mean. This 1-NNclassifier has the capability of combining together features belonging to different sets of data. Also the theoretical promise provided claiming that the stochastic discrimination has the upper hold of clubbing together the max and average voting merits. In a partial respect, this work derives the decision tree based C4.5 algorithm. Multiple classes are trained with the posteriori probability and the probabilities thus obtained outperforms the production rule manifesto that is been decreasing throughout the learning curve of the boosting mechanism.

**Chen and Ren (2021) [17]**, chose a ensemble based Gaussian regression models. The work mostly targets on complex computing intense processes. As the complexity increases, the skewed distribution of the original sampling techniques and bagging algorithms [18] are marked as well performing. In particular, the notion of linear reparability with soft-margin boosting shows proven and benchmarking less loss in problem marginalization around the feature concentric spot margins. Table -1 below provides the comparison of the various ensemble methods.

**Table 1:** description of popularly used ensemble mechanism

| Ensemble method | Behavior | Testing and additiveness | Content strength |
|---|---|---|---|
| Voting ensemble | Parallel | Not fitting | Abstractual content |
| Adaptive Contextual Buffers | Both Parallel and gated parallel | Fitting for the testing scenario | Confidence level of hidden content |
| associate Switches | Gated Parallel | Confidence level additiveness | Rank confidence |
| Differentiated increment | Cascaded parallel | Fitting for the confidence level | Rank confidence level |
| Product min-max ensemble | Hierarchical cascade | Not fitting for the additiveness | Abstractual context |

## III. EXPERIMENTAL SETUP

A dataset fed as input for any language model has a potentially unbounded, ordered sequence of data items with varying time intervals. A generalized dataset has the following characteristics which serve as the feature driven marginal. These include data time stamp, readiness of the data, plausibility of the data, conjunction of the data, susceptibility of the data and labeling cost of the data. Drifting mechanism [19] in the non-stationary datasets: Drift can be thought of as hidden contextual information that is unknown to the deployed algorithm. Drifts can be mitigated using the knowledge possessed from the previous classification strategies. A drift generated by a source at a specific point of time is defined as:

$D^t$ (x, y): t + delta(t), where "t" is the time stamp and delta(t) is the timing incremental. "x" is chosen such that $D^t$ (x, y) $\neq D^{t+delta(t)}$ (x, Y). This whole setup has to be susceptible with prior probabilities P(y) and class conditional probabilities P(x|y), the conditional probability. Drifts are often tagged along with the severity, decisive power and interval of the data generation. Gradual incremental [20] with the sampling aspect P [epoch + t], suffices the ensemble-variance trade-off. "Epoch" is the iterative cycle for the entire train-test ratio of 80:20. Sometimes datasets containing the blips and noise are categorized under the {rare event | outlier} boundary. The following figure 4 illustrates the restoration time compared with the performance of drift concept over the drift appearance pertained by every common dataset. The triggering updating and retaining out models are showing the maximum performance. Drift detectors are often the mechanism for replicating the warning signals that are restored over time. The entire drift detection mechanism can be summarized as follows:

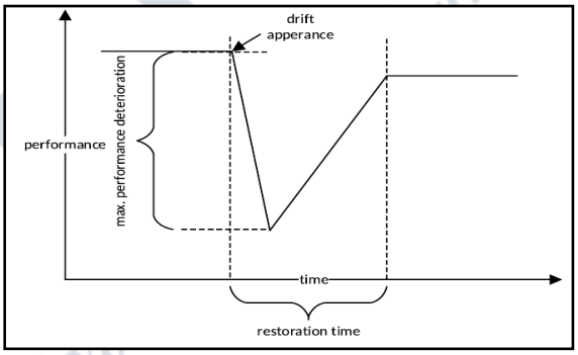



**Figure 4.** drift model restoration time (drift appearance)

1. Drifting through-put based on probabilistic models
2. Drifting outliers for the linear dataset pruning
3. Methods for window monitoring for the interval time upgradation
4. Ensemble-variance contextual incremental planning

Among these the unsupervised virtual drift detection mechanism has the upper hold of aligning with the statistical tests having the capability to handle the non-parametric drifts. In a nutshell, it is observed that a miniature variation in the ensemble detector shows the major deviation in the selective decision making scenario.

## IV. ANALYSIS OF ENSEMBLES WITH THE SUPERVISED REGRESSION METRICS:

Realizing the ensemble component [21] classifiers for the complete analysis involves multiple chunks of knowledge repositories. These repositories are many times unhelpful for non-stationary data streams. A typical chunk based ensemble [22] classification stabilizes the involuntary hikes occurring in the decision making stage. Chunks are often constructed based on the following criterion:

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Vol 12, Issue 05, May 2025

1. Every dataset miniscule $C_i$ ε U, where the $C_i$ is the ensemble component in the $i^{th}$ iteration and U is the dataset converted in to a data stream.
2. Whenever a new candidate classifier enters the decisive boundary, $C_c$ is embedded and de-duplicated with the elevated U. Here $C_c$ is the embedded de-duplicated parameter.

There are few other alternate chunk based approaches exists claiming to be the prediction accuracy improvisers. Often these alternates ends-up with over-fitted decisive boundary manipulations missing the small variations. Ignoring these small variations sometimes through huge drifts in the intermediary datasets.

## V. OUTLIER STATISTICS

Considering the dataset perturbation, an ensemble size of increasing timing interval outliers, the analysis error is compared with the candidate dimension and the probability prediction ratio. The difference between actual value and the "void predictability" is plotted. For a fitting ensemble validation and idealization the false/hit rate is also considered while plotting the analysis. The figure 5 illustrates the ensemble deviations with the impact of ensemble false positive normalized values.

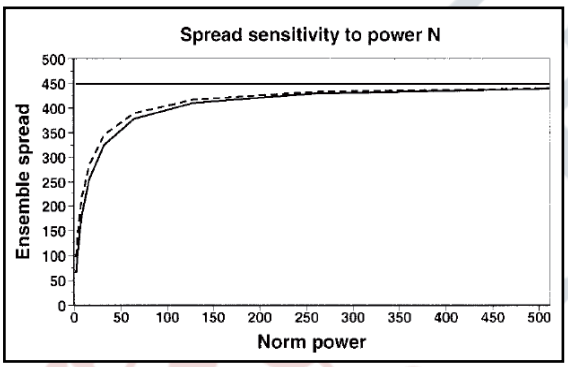



**Figure 5.** Ensemble deviation with positive normalized values

## VI. CONCLUSION

The initial section of this article, stationary and non-stationary characteristics of the ensemble and drifting mechanisms are examined. The detailed survey of ensemble with drifting concept is made summarily to project the class imbalance nature of the dataset even after the initial trimming and pruning. As mentioned in this article, higher the relationship among the fields of datasets, higher the predictive performance and components diversity. Especially in the re-occurring drifts, non-severe drifts and drift outliers. In recent days, it is recorded in the literature that self-ensembles due to hectic inclusion and non-transparency in the neural network layers, sure to produce wobbling prediction accuracy. This could probably raise the need for the reverse-tuned ensemble feedback mechanism.

## REFERENCES

[1] Ya-Lin Zhang, Qitao Shi, Meng Li, Xinxing Yang, Longfei Li, Jun Zhou, A Classification Based Ensemble Pruning Framework with Multi-metric Consideration, Intelligent Systems and Applications, 10.1007/978-3-030-82193-7_44, (650-667), (2022).

[2] Cai, Yu, Saugata Ghose, Erich F. Haratsch, Yixin Luo, and Onur Mutlu. "Error characterization, mitigation, and recovery in flash-memory-based solid-state drives." Proceedings of the IEEE 105, no. 9 (2017): 1666-1704.

[3] Smit, R. H. M., Carlos Untiedt, and J. M. Van Ruitenbeek. "The high-bias stability of monatomic chains." Nanotechnology 15.7 (2004): S472.

[4] Park SB, Park DH. The effect of low-versus high-variance in product reviews on product evaluation. Psychology & Marketing. 2013 Jul;30(7):543-54.

[5] Briscoe, Erica, and Jacob Feldman. "Conceptual complexity and the bias/variance tradeoff." Cognition 118.1 (2011): 2-16.

[6] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[7] Aakash, V., et al. "Forecasting of Novel Corona Virus Disease (Covid-19) Using LSTM and XG Boosting Algorithms." Data Analytics in Bioinformatics: A Machine Learning Perspective (2021): 293-311.

[8] Fan, Junliang, et al. "Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data." Agricultural water management 225 (2019): 105758.

[9] Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support." arXiv preprint arXiv:1810.11363 (2018).

[10] Hansen, Lars Kai, and Peter Salamon. "Neural network ensembles." IEEE transactions on pattern analysis and machine intelligence 12.10 (1990): 993-1001.

[11] Everett D, Ke W, Paquet JF, Vujanovic G, Bass SA, Du L, Gale C, Heffernan M, Heinz U, Liyanage D, Bayesian constraints on the transport coefficients of QCD matter. Physical Review C. 2021 May 14;103(5):054904.

[12] Bilenko, Misha. Scaling up decision tree ensembles. Technical report. http://hunch. net/~ large_scale_survey/ TreeEnsembles. pdf, 2011.

[13] Perrone, Michael P., and Leon N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. Brown Univ Providence Ri Inst for Brain and Neural Systems, 1992.

[14] Tumer, Kagan, and Joydeep Ghosh. "Error correlation and error reduction in ensemble classifiers." Connection science 8.3-4 (1996): 385-404.

[15] Griffin, Matthew, Susanne Herrmann, and Friedrich A. Kittler. "Technologies of writing: interview with friedrich a. Kittler." New Literary History 27.4 (1996): 731-742.

[16] Jain, Rajendra K., Dah-Ming W. Chiu, and William R. Hawe. "A quantitative measure of fairness and discrimination." Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA 21 (1984).

[17] Liu, W., Yan, X., Chen, G. and Ren, Z., 2012. Recent advances in thermoelectric nanocomposites. Nano Energy,

1(1), pp.42-56.
[18] Bauer, Eric, and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." Machine learning 36.1 (1999): 105-139.
[19] Cavanagh, Patrick, and Stuart M. Anstis. "Brightness shift in drifting ramp gratings isolates a transient mechanism." Vision Research 26.6 (1986): 899-908.
[20] Duan, Feng, and Lili Dai. "Recognizing the gradual changes in sEMG characteristics." IEEE Transactions on Industrial Electronics 64.5 (2016): 4276-4286.
[21] Laubach, Mark, Marshall Shuler, and Miguel AL Nicolelis. "Independent component analyses for quantifying neuronal ensemble interactions." Journal of neuroscience methods 94.1 (1999): 141-154.