10.31341/jios.47.1.9
Open Access
UDC 004.8
Original Scientific Paper

Data-Centric Optimization Approach for Small, Imbalanced Datasets

Vladislav Tanov *vlad.k.tanov@gmail.com*
Faculty of Economics and Business Administration
Sofia University St. Klimenr Ohridski, Sofia, Bulgaria

Abstract

Data-centric is a newly explored concept, where the attention is given to data optimization methodologies and techniques to improve model performance, rather than focusing on machine learning models and hyperparameter tunning. This paper suggests an effective data optimization methodology for optimizing imbalanced small datasets that improves machine learning model performance.

This paper is focused on providing an effective solution when the number of observations is not enough to construct a machine learning model with high values of the estimated magnitudes. For example, the majority of the observations are labeled as one class (majority class), and the rest as the other, commonly considered as the class of interest (minority class). The proposed methodology does not depend on the applied classification models, rather it is based on the properties of the data resampling approach to systematically enhance and optimize the training dataset. The paper examines numerical experiments applying the data centric optimization methodology, and compares with previously obtained results by other authors.

Keywords: imbalanced dataset, classification, data centric, optimization, machine learning, artificial intelligence.

1. Introduction

Highly imbalanced data appears in many real-world domains, such as 1) detection of cardiovascular (heart) and liver diseases, diabetes, 2) detection of oil spills in satellite images, 3) information retrieval and filtering tasks and so on [1], [2]. The task of improving the effectiveness of a class-imbalance problems is very important aspect for averting a threat before it occurs, effectively applying medical treatments etc.

Many authors, working in this direction, proposed different approaches addressing data and algorithmic methodologies. For example, some data methodologies [3] are related to the use of resampling, bootstrapping and feature selection. The application of resampling methods includes neural networks [4], ensemble models [5], time series [6] and others. Resampling procedures for SVM hyper parameter selection are investigated by Wainer [7]. Authors have considered resampling procedures such as an estimate performance based on repeatedly dividing a given set into training and test set.

Azbeg and co-authors have considered bootstrapping, randomly drawing without replacement, to construct a collection of estimators, also known as Random Forest (RF), to improve the diabetes classification algorithmic methodology. Azbeg compared results with different algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree-based (DT), Adaptive Boosting (ADABoost), Artificial Neural Network (ANN), Logistic Regression (LR), Deep Learning (DL), and concludes it can improve the accuracy of the diabetes classification problems [8].

Singh and co-authors investigated feature selection technique in attempt to evaluate liver diseases and improve classification algorithmic methodology. Feature selection is used for removal of redundant and irrelevant data to increase learning accuracy and reduce noise. The research continues to compare results of both with and without selection technique on different classifiers with resampling procedure, also known as cross validation (CV). Six classifiers have been considered, as follows: LR, Naïve-Bayes (NB), Java implementation of DT C4.5 (J48), Sequential Minimum Optimization (SMO), Weka Abstract Classifier (IBk) and RF [9].

Bohacik and Zabovsky studied a probabilistic implementation with given supervised discretization, utilizing heart disease expert domain knowledge [10]. The algorithmic methodology was applied in Waikato Environment for Knowledge Analysis as class NaïveBayes with Fayyad-Irani's discretization of numerical attributes [11]. The research is based on k-fold (k=10) cross-validation, and uses sensitivity, specificity and their sum as measures. Sensitivity (true positive rates) represents the ability of the algorithm to identify true positive (*TP*) cases in respect to all positive outcomes, with following expression: $\frac{TP}{TP+FN}$. False negative (FN) would be considered as negative cases while actually being positive. Specificity states the ability of the algorithm to identify true negative rates (*TN*) cases in respect to all negative outcomes, as follows: $\frac{TN}{TN+FP}$. False positive (*FP*) would be considered as positive cases while actually being negative. Bohacik and Zabovsky use the sum of sensitivity and specificity in [10] as an overall point system to benchmark algorithms.

This paper's intention is to apply a new effective approach centered around the data, also referred as data centric optimization approach, to improve machine learning models (classifiers) [12]. The algorithm resamples from the given data to produce an optimized balanced data sub-set. An optimized balanced data sub-set is derived based on an objective function that minimizes model's error (false positive and false negative error). To ensure proper result comparisons, this methodology is benchmarked against previously discussed approaches and various algorithms using the same datasets, described below in Table 1.

Datasets	Variables	Classes/Observations	Source
Diabetes	8	Class 0 – 500 Class 1 - 268	https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f
Heart	13	Class 1 – 150 Class 2 - 120	http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
Indian Liver Patient Dataset (ILPD)	10	Class 1 – 416 Class 2 - 167	https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

Table 1. Public datasets and sources.

During the optimization phase, the algorithm randomly resamples unique data sub-set from the given data, assuring under-sampling of the majority class and producing a balanced, unique data sub-set. Then data sub-set is divided into train and validation set, according to the classical split of 80% and 20%, respectively, and fitted to a classifier model.

The optimized, balanced data sub-sets (train and validation) are derived and stored based on an objective function, which minimizes model's error, noted as model error rate and stored each iteration for comparison. This process repeats until minimization is reached and/or random sampling is exhausted (ranging between 0 and 100).

Next, during the test (classification) phase, the optimized, balanced data sub-sets will be fitted into a newly instantiated classifier for final model test evaluation. The proposed algorithm is tested in (RAM: 16GM, CPU: 2.6GHz 6-Core Intel Core i7) and illustrated below:

Let:

- $i \in \{0 \ldots 100\}$
- r_i – random under-sampling integer
- n – length of the given dataset
- m – minority class length
- R – list of integers
- D_n – given dataset
- X_n – random variable (# of variables in D_n)
- Y_n – response variable (# of classes in D_n), Y = 1,…K, where K>=2
- $D_{i,m}$ – balanced, under-sampled data sub-set:
  $D_{i,m} = ([X_1,Y_1],\ldots.[X_n,Y_n] \mid r_i , m)$, where [X,Y] is independent of D_n
- *mer* - model error rate, *mer* = 100
- T_o and V_o – optimized train and validation sub-sets

ALGORITHM: *OPTIMIZATION PHASE*

Initialization of variables listed above.
set ***optimized*** *= False*
while *not* ***optimized***
    draw random integer – r_i
    if random integer (r_i*) not in list of integers (R)*
        append random integer (r_i*) to R*
        $\boldsymbol{D_{i,m}}$ *= undersample(*$D_n \mid r_i , m$*)*
        split $D_{i,m}$ *into train and validation sets (80/20):*
        $\boldsymbol{T_i, V_i}$ *= train_test_split (*$D_{i,m} \mid .20$*)*
        $\boldsymbol{C_i}$ *= build classifier*
        fit train set to C_i *and calculate false positive error (FPE) and false negative error (FNE). Keep track of* C_i *error:*
        $\boldsymbol{error_{Ci}} = C_i\ (T_i,\ V_i) = \Sigma\ [(FPE_i \mid C_i,\ D_{i,m}),(FNE_i \mid C_i,\ D_{i,m})]$
        evaluate current model error rate (mer)
        if ***mer*** *is greater than* $error_{Ci}$
            $\boldsymbol{mer} = error_{Ci}$
            $\boldsymbol{T_o} = T_i$
            $\boldsymbol{V_o} = V_i$
    if lenght of R is greater than 100: ***optimized*** *= True*
end

The idea of optimizing a balanced dataset (minimizing model's error rate) is based on a filtering technique, also known as information gain. Shaltout and co-authors studied information gain (IG) as a feature selection method for the efficient classification of influenza. In their research IG was used to identify the feature(s) possessing the most information, based on a specific class, and was derived from entropy, as entropy is inversely proportional to IG [13],[14]. In this research, minimizing model's error, thru under-sampling of the majority class, plays the role of information gain, i.e., selecting a majority class sub-set possessing the most information. In other words, the data centric optimization approach filters out "bad" and/or "noisy" majority class datapoints from the dataset to produce balanced, optimized sub-sets (training and validation).

2. Numerical Experiments

2.1. Hypothesis

Data centric optimization approach is not effective solving classifications tasks on imbalanced datasets with "small" number of observations. To prove the hypothesis, this paper includes experiments and comparison with methodologies described in the

introductions. Moving forward, experiments and comparisons are combined into three sections, as follows: bootstrapping without replacement, feature selection, and probabilistic implementation with given supervised discretization.

2.2. Bootstrapping with replacement

Diabetes is a common disease and patients living with such illness have to continuously self-control it to avoid fatal consequences. Usual cause of diabetes is high level of glucose (blood sugar). Azberg and co-authors conducted research in an attempt to build a predictive model, Random Forest (RF), to make prediction estimates for patients with diabetes. RF has wide applicability, and along with its ease of use and good performance is considered as a standard method in supervised machine learning. Azberg and co-authors' algorithm (ACA) consists of aggregation a collection of estimators constructed from bootstrap samples with replacement in the training set, also called random weak learners. A RF is an aggregation of randomly generated weak learners and defines a prediction rule that corresponds to the majority vote in classification [8]. The accuracy given from their experiments and comparison is listed in Table 2. Class 0 denotes cases when patients have diabetes.

Datasets	Classes/Observations	ACA RF Accuracy	**Our Proposition RF Accuracy**
Diabetes 1	Class 0 – 500 Class 1 – 268	78.65	**87.04**
Diabetes 2	Class 0 – 1316 Class 1 – 684	99.5	**99.65**
Diabetes 3 (1 and 2 combined)	Class 0 – 1816 Class 1 – 952	99.8	**100**

Table 2. Results and comparison between ACA [8], and data centric optimization approach.

As seen in the table above, ACA underperforms in comparison with our proposed algorithm, and as the number of observations increases, the accuracy of the RF model increases as well. This is an expected performance, although it is the main idea behind the ACA. The major difference between ACA and our proposition lies in the biasness of the RF model.

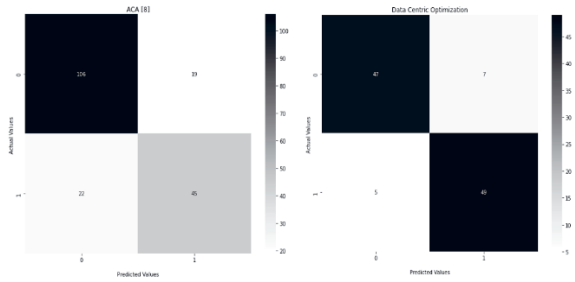

Figure 1. Confusion matrix results comparison for Diabetes Dataset 1.

Looking deeper and analyzing the results from Figure 1, confusion matrix result comparison of Dataset 1, we observe, due to the imbalanced data, that ACA is biased towards cases when diabetes is presents. In other words, ACA's ability of the algorithm to identify true negative cases, also known as specificity, is 70.31.

This implies that, on average, the algorithm assigns diabetes treatment to patients with no diabetes ~30% of the time. In contrast, the specificity of our proposed algorithm is 87.5, and it will mistakenly assign diabetes treatment, on average, only 12.5 of the time (more than half lesser than ACA).

2.3. Feature Selection

According to Centers of Disease Control and Prevention (CDC), there is 4.5M (1.8%) adults diagnosed with liver disease in the US (with a death rate of 15.7 per 100K) [15]. Singh and co-authors investigated feature selection technique in attempt to evaluate liver diseases and help improve patient treatment.

The research continues to compare results of both with and without feature selection with resampling procedure, considering six classification algorithmic methodologies, as follows: Logistic Regression (LR), Naïve-Bayes, Java implementation of DT C4.5 (J48), Sequential Minimum Optimization (SMO), Weka Abstract Classifier (IBk) and Random Forest (RF) [9].

Indian Liver Patient dataset (ILPD) from the UCI Repository [16] has been used in this research, containing 416 liver (class 1) and 167 none liver patient (class 2) records. Correlation-based feature selection evaluator and Greedy Stepwise were applied, as feature selection and search method with 10-fold cross validation, in WEKA tool [17] resulting in 5 features: Total Bilirubin (TB), Direct Bilirubin (DB), Alkaline Phosphatase (Alkphos), Alamine Aminotransferase (Sgpt) and Asparate Aminotransferase (Sgot) in [9]. Accuracy comparison shown below in Figure 2.

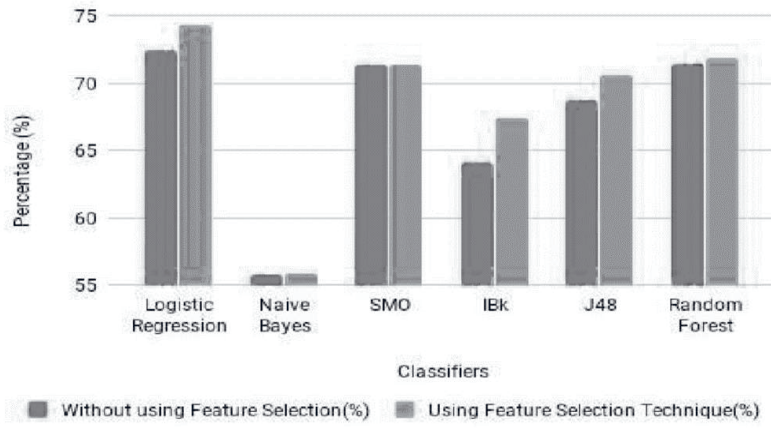

Figure 2. Result comparisons between different classifiers on basis of correctly classified instances [9].

As seen above in Figure 2, LR achieved the highest accuracy with feature selection (FS) or without feature selection technique of 74.36 and 72.5, respectively. Random Forest performance was placed second with 71.87 and 71.53.

This shows that feature selection technique underperforms in comparison with our proposed algorithm. Our proposed algorithm, data centric optimization approach using Random Forest, achieved significantly higher Test Accuracy of 92.54, as seen below in Table 3. In addition, due to accuracy being misleading and not a preferred metric in cases with imbalanced small datasets, we are reporting our data centric optimization approach test results for precision, recall, AUC and F1- Score in Table 4.

Datasets	Classes / Observations	Singh and co-authors LR Accuracy		Singh and co-authors RF Accuracy		**Our Proposition RF Accuracy**
		No FS	FS	No FS	FS	
Indian Liver Patient dataset (ILPD)	Class 1 – 416 Class 2 – 167	72.5	74.36	71.53	71.87	**92.54**

Table 3. Results and comparison between Singh and co-authors [9], and data centric optimization approach.

Indian Liver Patient dataset (ILPD)	Our Proposition Test Metric Results AUC = 92.5			
	Accuracy	Precision	Recall	F1
Class 1 (34)	92.54	91.43	94.12	92.75
Class 2 (33)	92.54	93.75	90.91	92.31

Table 4. ILPD Test results for data centric optimization approach, using Random Forest.

2.4. Probabilistic implementation with given supervised discretization.

Bohaick and Zabovsky studied a probabilistic implementation in an attempt to detect heart disease cases. CDC relates the term “heart disease” to several types of heart conditions (most common is coronary artery disease), which is the leading cause of death in United States [18]. Bohacik and Zabovsky research is based on UCI Repository’s Statlog Heart dataset, containing 120 heart (class 2) and 150 none heart disease (class 1) records [19].

Their research employs Equal Frequency Discretization algorithm, with focus on refining preprocessing methods. The research presents improvements of achieved accuracy with added discretization, selected based on the smallest average entropy. Research results and comparisons between the proposed (NB-Mod) and additional

algorithms are derived from the summation of Sensitivity and Specificity, listed in Table 3 below.

Algorithm	Sensitivity	Specificity	Sum
NB-Mod	0.900	0.842	1.742
NB	0.840	0.817	1.657
MLP	0.880	0.800	1.680
DT	0.840	0.692	1.532
NN	0.773	0.717	1.490
Our Proposition of RF	**0.96**	**1.00**	**1.96**

Table 5. Experimental results of Bohacik and Zabovsky approach (NB-Mod) and comparison with other machine learning algorithms in [10].

As seen in Table 4 above, Bohaick and Zabovsky approach achieved the highest Sensitivity and Specificity of .90 and .842, respectively, in comparison with other machine learning algorithms. Our proposed algorithm, data centric optimization approach, outperforms significantly their model reaching Test Sensitivity and Specificity of .96 and 1.0 (Sum of 1.96). In addition, we are reporting our data centric optimization approach test results for precision, recall, AUC and F1- Score in Table 5.

Heart	Our Proposition Test Metric Results AUC = 97.9			
	Accuracy	Precision	Recall	F1
Class 1 (24)	97.9	1.00	95.83	97.87
Class 2 (24)	97.9	96.0	1.00	97.96

Table 6. Satlog Heart Test results for data centric optimization approach, using Random Forest.

3. Conclusion

This research work presented a new approach centered around the data, referred as data centric optimization. Three experiments were conducted using the following datasets: diabetes, liver and heart disease, listed in the Introduction section under Table 1. Research results were tested and compared with various approaches utilizing different classifiers.

Evidently, data centric optimization approach outperformed bootstrapping without replacement (Section 3.2), feature selection (Section 3.3) and probabilistic implementation with given supervised discretization (Section 3.4). This gives strong attention to data centric optimization approach and its effectiveness solving classification tasks on imbalanced datasets with “small”.

Acknowledgements

This work is supported by the project under a Sofia University St. Kliment Ohridski grant for 2023 year.

The author would like to extend his gratitude to Prof. Ivan Ganchev Ivanov for the idea generation and support during the implementation of the data centric optimization methodology, as part of the completion of this thesis under his supervision.

References

[1] B. Raskutti and A. Kowalczyk, "Extreme rebalancing for svms: a case study". *SIGKDD Explorations*, 2004. Available at: https://storm.cis.fordham.edu/~gweiss/selected-papers/extreme-rebalancing-svms.pdf [Accessed: August 25, 2022].

[2] G. Wu and E. Chang. Lecture, Topic: "Class-Boundary Alignment for Imbalanced Dataset Learning". *ICML Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, 2003. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.9007&rep=rep1&type=pdf [Accessed: August 25, 2022].

[3] W. Satriaji, R. Kusumaningrum, "Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis." *2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2018. 10.1109/ICICOS.2018.8621648 9 [Accessed: March 01, 2023].

[4] L.S. Chen and S.J. Cai, "Neural-network-based resampling method for detecting diabetes mellitus", *SpringerLink*, November 2015. Available at: https://link.springer.com/article/10.1007/s40846-015-0093-9 [Accessed: August 25, 2022].

[5] E. Snieder, K. Abogadil and U.T. Khan, "Resampling and ensemble techniques for improving ann-based high-flow forecast accuracy," *Hydrology and Earth System Sciences*, May 2021. Available at: https://hess.copernicus.org/articles/25/2543/2021/ [Accessed: August 25, 2022].

[6] N. Moniz, P. Branco and L. Torgo, "Resampling strategies for imbalanced time series forecasting", *SpringerLink*, February 2017. Available at: https://link.springer.com/article/10.1007/s41060-017-0044-3 [Accessed: August 25, 2022].

[7] J. Wainer and G. Cawley, "Empirical evaluation of resampling procedures for optimising SVM hyperparameters", *Journal of Machine Learning Research*, vol. 18, February 2017. Available at:

https://jmlr.org/papers/volume18/16-174/16-174.pdf [Accessed: August 25, 2022].

[8] K. Azbeg, M. Boudhane and O. Ouchetto, "Diabetes emergency cases identification based on a statistical predictive model", *Journal of Big Data*, March 2022. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00582-7 [Accessed: August 25, 2022].

[9] J. Singh, S. Bagga and R. Kaur, "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques", *Procedia Computer Science*, March 2020. Available at: https://www.sciencedirect.com/science/article/pii/S187705092030692X [Accessed: August 25, 2022].

[10] J. Bohacik and M. Zabovsky. (2019). *Discretization for Naive Bayes Taking the Specifics of Heart Data into Account*. Journal of International and Organizational Science, vol. 43, no. 1. Retrieved August 25, 2022, from: https://jios.foi.hr/index.php/jios/article/view/1210

[11] U. M. Fayyad and K. B. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," in International Joint Conference on Uncertainty in AI, 1993, pp. 1022-1027. Available: https://trs.jpl.nasa.gov/bitstream/handle/2014/35171/93-0738.pdf?sequence=1&isAllowed=y [Accessed: August 25, 2022].

[12] C. Bartel, "Data-centric approach to improve machine learning models for inorganic materials", *Cell Press*, November 2021. Available at: https://www.cell.com/patterns/pdf/S2666-3899(21)00249-X.pdf [Accessed: August 25, 2022].

[13] N. Shaltout, M. Elhefnawi, A. Rafea and A. Moustafa, "Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts", *Engineering and Computer Science*, July 2014. Available at: https://www.researchgate.net/publication/286743906_Information_Gain_as_a_Feature_Selection_Method_for_the_Efficient_Classification_of_Influenza_Based_on_Viral_Hosts [Accessed: August 30, 2022].

[14] K. Leung, K. H. Lee, J. Wang, E. Ng, H. Chan, S. Tsui, et al., "Data mining on DNA sequences of Hepatitis B virus", *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, March 2011, Available at: https://www.researchgate.net/publication/49754043_Data_mining_on_DNA_sequences_of_Hepatitis_B_virus [Accessed: August 30, 2022].

[15] FastStats. *FastStats - Chronic Liver Disease or Cirrhosis*, www.cdc.gov, January 2022, Available at: https://www.cdc.gov/nchs/fastats/liver-disease.htm [Accessed: 24 August 2022].

[16] UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set. *UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set*, archive.ics.uci.edu, archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) [Accessed: 24 August 2022].

[17] H. Written, E. Frank, A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA: Morgan Kaufmann, 2016. [E-book] Available at: Elsevier [Accessed: 24 August 2022].

[18] CDC. "Heart Disease | Cdc.Gov." *Centers for Disease Control and Prevention*, www.cdc.gov, June 2022, Available: https://www.cdc.gov/heartdisease/index.htm#:~:text=Heart%20disease%20is%20the%20leading,can%20lead%20to%20heart%20attack [Accessed: 24 August 2022].

[19] UCI Machine Learning Repository: Statlog (Heart) Data Set. *UCI Machine Learning Repository: Statlog (Heart) Data Set*, archive.ics.uci.edu, archive.ics.uci.edu/ml/datasets/statlog+(heart). [Accessed: 24 August 2022].