Improved bio-inspired technique for big data analytics and machine learning speed optimization

Andronicus A. Akinyelu[1][0000-0003-2172-0755]

[1] Department of Computer Science and Informatics, University of the Free State, Bloemfontein, Free State, South Africa

Abstract

Big Data Analytics (BDA) is progressively becoming a popular practice implemented by many organizations, because of their potential to discover valuable in-sights for improved decision-making. The International Data Corporation predicts that the Global Datasphere will grow from 33 Zettabytes in 2018 to 175 Zettabytes in 2025. Obviously, we are currently in the era of Big Data (BD), and the rate of data growth is very alarming. Unfortunately, BD does not offer a lot of value in its unprocessed form. Therefore, to unlock the great potentials of BD, we need efficient BDA methods. Machine Learning (ML) algorithms are one of the most efficient tools suitable for data analytics, however, some ML algorithms cannot effectively handle BD; their computational complexity increases with in-crease in data size. Therefore, some researchers introduced various techniques for improving the speed of ML algorithms, including feature selection techniques, in-stance selection techniques, sampling, and distributed computing. However, most of them failed to achieve a balanced trade-off between storage reduction and predictive accuracy [1]. Therefore, this paper introduces a boundary detection and instance selection technique for improving the speed of ML-based BDA, called Ant Colony Optimization Instance Selection Algorithm for Machine Learning (ACOISA_ML)). The key highlights of ACOISA_ML are outlined below:

Boundary identification: The first stage of ACOISA_ML is the boundary identification stage. Unlike other ACO-based instance selection techniques that directly use ACO algorithm for instance or feature selection, ACOISA_ML use ACO algorithm for boundary identification. It adopts the concept of ACO edge selection to search for different boundaries (not to select instances). To the best of author's knowledge, this study is one of the first studies that adopt the concept of ACO edge detection for instance selection problems. This concept is mostly used for image edge detection (and not data boundary identification) [2-4].

Boundary instance selection: The second stage of the proposed technique is the boundary instance selection stage. After identifying different boundaries, ACOISA selects the best boundary and use k-NN to select the relevant instances for training (that is, instances close to the best-identified boundary).

Heuristic value computation: This study introduces a novel method for computing heuristic value for ACO. This method is suitable for boundary instance se-lection problems. ACOISA_ML is designed to use the proposed computation method to cal-

culate the heuristic value for each instance in the dataset. As afore-mentioned, ACO is used to identify the best boundary instance, that is, the in-stance with the highest pheromone value. Hence, the heuristic value for each in-stance is designed to reflect the boundary information for each instance.

The technique was evaluated on five ML algorithms, namely: (Artificial Neural Network (ANN), Random Forest (RF), Naïve Bayes (NB), k Nearest Neighbor (k-NN), and Logistic Regression (LR)). In this study, we refer to the models produced by the full dataset as standard models, and the models produced by the reduced subset as hybrid models. Finally, we compare the hybrid models to the standard models based on the following criteria: (i) the ability to preserve prediction accuracy (ii) training speed (iii) storage reduction percentage, and (iv) algorithm time (or instance selection time). All the datasets used in this study were obtained from the UCI data repository [5]. Annexures 1 and 2 shows the average training speed and prediction accuracy produced by the standard models (denoted as Standard) and hybrid models (de-noted as Hybrid). As shown in the Annexures, the hybrid models achieved better training speed than the standard models without significantly affecting their pre-diction accuracy. Moreover, the right-hand side of Annexure 1 shows the average algorithm time (denoted as Alg-T) and the average storage reduction percentage (denoted as Av-Sto) achieved by ACOISA_ML. The storage reduction percentage represents the fraction of instances selected after data reduction. As shown in the Annexure, ACOISA_ML reduced the storage size of the evaluated big datasets by over 55% (in most cases) without substantially affecting their quality. Moreover, ACOISA_ML achieved good instance selection time. It used an average of 36.8 seconds to reduce the largest dataset evaluated in this study (i.e. Twitter dataset). This shows the effectiveness of ACOISA_ML for BDA.

In addition, ACOISA_ML was compared to four recent instance selection algorithms, namely: LDIS, LSSM, LSBO, and ISDSP. The algorithms were evaluated on SVM, hence we first evaluated ACOISA_ML on SVM before comparing it to the algorithms. Annexure 3 shows the prediction accuracy (denoted as accuracy) and storage reduction percentage (denoted as storage) for the four algorithms. The best prediction accuracy for each dataset is underlined. As shown, ACOISA_ML outperformed LSSM in prediction accuracy in 6 out of 11 datasets and outperformed LSBO in 7 out of 11 datasets. Moreover, the results show that ACOISA_ML outperformed LDIS and ISDSP in 9 out of 11 datasets. Furthermore, T-test statistical analysis was performed to evaluate the speed improving capacity of ACOISA_ML. Specifically, we compared the training speed produced by the hybrid models of ANN and RF to the training speed produced by the standard algorithms. The P-values produced by all the test analysis are less than 0.05, hence we can conclude with 95% confidence level that ACOISA_ML is significantly faster, in terms of training speed, than the analyzed standard algorithms. Overall, the results show that the proposed technique is suit-able for fast and simplified BDA and ML speed optimization.

Keywords: Big data analytics, Machine learning, Instance selection, Data reduction, Speed optimization.

References

1. E. Leyva, A. González, and R. Pérez, "Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective," Pattern Recognition, vol. 48, no. 4, pp. 1523-1537.
2. J. Tian, W. Yu, and S. Xie, "An ant colony optimization algorithm for image edge detection," in 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence), 2008, pp. 751-756.
3. M. Nayak and P. Dash, "Edge Detection Improvement by Ant Colony Optimization Compared to Traditional Methods on Brain MRI Image," Communications on Applied Electronics (CAE), vol. 5, no. 8, pp. 19-23, 2016.
4. A. Gautam and M. Biswas, "Edge Detection Technique Using ACO with PSO for Noisy Image," Singapore, 2019, pp. 383-396.
5. K. Bache and M. Lichman. (2013), "UCI machine learning repository". available at: http://archive.ics.uci.edu/ml (accessed 12-May-2017).

Annexures

Annexure 1. Average training time for the hybrid model and standard model

Datasets	KNN		ANN		RF		LR		NB		ACOISA_ML	
	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid	Sel-T	Av.Sto
Landstat	0	0.001	115.16	40.811	4.82	8.458	4.79	1.212	0.13	0.036	6.7908	31.567
Letter	0.02	0.047	365.31	177.991	174.53	77.836	11.29	7.671	0.1	0.058	101.42	43.75
Mushroom	0.01	0.001	79.44	19.277	1.24	0.211	1.87	0.458	0.12	0.025	14.758	35.435
Optdigit	0.03	0.004	278.16	157.821	26.64	18.089	2.6	2.289	0.08	0.05	32.614	52.315
Page-bloc	0.02	0.004	25.61	10.482	3.02	1.793	4.16	1.243	0.04	0.014	19.839	45.678
Shuttle	0.04	0.017	255.04	109.026	1757.58	27.482	18.75	7.617	0.3	0.077	645.186	43.678
Twitter	0.06	0.015	8859.44	485.939	69.84	2.902	275.06	6.987	4.46	0.208	503.838	6.219
USPS	0.03	0.001	5047.81	2420.235	509.16	227.328	19.06	8.839	0.65	0.319	97.992	43.89
Pentdigit	0.02	0.002	74.25	32.38	108.86	39.143	4.46	1.6	0.07	0.027	15.175	33.36
Waveform	0	0	50.33	11.299	1.1	0.206	6.29	1.178	0.08	0.011	4.056	24

Key: Standard: time produced by the standard algorithms, Hybrid: time produced by the hybrid model, Sel-T: average instance selection time (in seconds), Av-Sto: Average storage percentage

Annexure 2. Average prediction accuracy for the hybrid and standard model

Datasets	KNN		ANN		RF		LR		NB	
	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid	Standard	Hybrid
Landstat	90.55	86.86	88.5	85.165	83.75	82.1	91.05	88.085	79.6	78.705
Letter	95.725	90.995	80.975	79.39	77.375	75.9675	96.175	91.9125	62.3	62.2725
Mushroom	100	99.91	98.966	99.015	95.4825	99.97	100	99.95	90.8296	91.945
Optdigit	97.8297	93.7841	96.5498	93.4613	92.3205	86.9004	97.3845	90.384	89.4268	83.9232
Page-bloc	96.0168	96.916	96.2361	97.328	96.4553	97.408	97.5333	97.948	90.846	92.764
Shuttle	99.9103	99.7752	99.7517	99.7062	96.8345	96.76	99.9931	99.9028	92.2069	92.5297
Twitter	96.0911	94.6939	96.4109	94.7831	96.5566	95.3936	96.6881	95.1853	94.9611	93.2472
USPS	95.1171	93.7369	94.3199	93.2835	89.5366	86.871	93.3732	92.3119	76.7813	75.1022
Pentdigit	97.7416	92.8845	89.8228	89.1367	89.8228	89.1367	96.5981	90.7919	82.1326	81.6352
Waveform	80.24	81.8	83.84	85.2667	87.08	87.5167	85.24	85.5167	81.02	80.6083

Key: Standard: average prediction accuracy (%) produced by the standard algorithm, Hybrid: average prediction accuracy produced by the hybrid model

Annexure 3. Comparison between ACOISA_ML and LDIS, LSSM, LSBO, ISDSP (SVM Classifier)

Datasets	ACOISA_ML		LDIS		LSSM		LSBO		ISDSP	
	Accuracy	Storage	Accuracy	Storage	Accuracy	Storage	Accuracy	Storage	Accuracy	Storage
Cardiotocography	71.25	26.13	62	14	67	86	62	31	59	10
Ecoli	84.97	67.21	77	8	83	91	74	17	78	10
Heart-statlog	82.44	61.73	81	7	84	85	81	33	78	10
Ionosphere	92.51	31.75	84	9	88	96	45	19	86	10
Landsat	84.81	45.1	84	8	87	95	85	12	84	10
Letter	89.10	25	75	18	84	96	73	16	74	10
Optdigits	92.13	52.31	96	8	99	98	98	8	97	10
Page-blocks	95.12	20.31	94	13	94	97	92	4	91	10
Parkinson	80.95	58.31	82	17	87	89	82	13	85	10
Segment	94.63	14.22	89	18	90	90	90	18	87	10
Wine	94.94	65.03	94	12	97	89	96	25	93	10

Key: Accuracy: average prediction accuracy (%) produced by the hybrid models, Storage: average storage reduction percentage produced by the instance selection techniques