Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 911-*918*

Research Article

Comparative Study of Novel Machine Learning Algorithms on a Scalable Data Set

Satyajit S. Uparkar, Sachin D. Upadhye, Kaushik R. Roy, Vishnu V. Budati

Department of Computer Application, Shri Ramdeobaba College of Engineering and Management, Nagpur, India,
uparkarss@rknec.edu, upadhyesd@rknec.edu, roykr@rknec.edu, budativv@rknec.edu

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: Scalability is a component of the concept or function which is capable of handling and performing well under an enhanced or large dataset. Scope of scalability in this concern can be related to data, features or interaction. The main problems are to basically resolve on the scalability issues related to supervised machine learning algorithms. The main aim of this research study is to evaluate and examine scalability of supervised machine learning algorithms. We have taken two novel supervised algorithms- L1 supervised algorithm and semi L1 supervised algorithm and applied on a massive data set of Brain tumor. Lastly the analysis based on various metrics viz. performance, accuracy, F1 score, recall and confusion matrix of the dataset with the two supervised machine learning algorithms are examined. The end results match with previous study even if on the medical domain massive data set, used in this research study.

Keywords: Scalability, L1 supervised algorithm, Semi L1 supervised algorithm, F1 score

1. Introduction

Scaling is a method to optimize the independent features included in the dataset in a specified range. Machine Learning (ML) is a broad approach in information systems, psychology, mathematics, prediction, neurobiology and several other disciplines. With ML, the issues can be overcome simply by creating a model that is a true reflection of the specified data set. ML has been a large structure from computer science to human brain imitation, which has introduced the field of statistics to a result derived that develops important mathematical numerical. In ML, most scalability issues are related with huge datasets, having millions of instances and/or say having bundle of features. ML is all about building processes that help your machine to learn. Learning is a method for identifying useful patterns or other data trends. ML algorithms are developed in order to be able to reflect a specific way of learning a problem. These algorithms may also provide indication of the relative complexity of learning in various settings. In these times, in the field of big data, machine learning is not like in the past machine learning.

The machine learning algorithms can be categorized in supervised and unsupervised types. Semi-supervised is the combination of these two types. A usual, machine learning process consists of feeding the data via the input pipeline based on ETL tools, doing a forward pass, computing loss, and then correcting the parameters with an objective to minimize the loss. The classification problem is one of the standard formulations of the supervised learning task: the learner is expected to learn a function that maps a vector to one of several classes by looking at several input and outputs

2. Literature Survey

(Markus G. et al.2015) analyzed a massive data set of earth data science on two perceptions of data volume and its dimensionality. Parallel and Scalable approach using DBSCAN and SVM techniques were used for the implementation. DBSCAN was used to detect the outliers and to reduce noise. SVM served classification for multispectral satellite images [2].

(Gupta et.al. 2016)has provided summary of big data analytics, with a special emphasis on data-intensive distributed machine-learning algorithms for big data. This study also carries out a comparative study on optimization and efficiency metrics of the parallel machine-learning data-intensively [3].

(Faraz Faghri et al. 2017) surveyed the existing scalable data mining and machine learning algorithms for bioinformatics data set. They have reflected the comparative analysis of each algorithm along with their challenges, pro and cons. The study offers the researchers to explore on the potential scalable alternatives applied in the versatile domain of computer science [4].

(Quan Zou et al. 2017) surveyed several emerging topics of scalable data mining techniques and applications for biomedicine or bioinformatics [5].

(Alam M.M. et al. 2018) analyze that supervised learning algorithm used to predict employee turnover on of the major issues of facing any organization. Eventually, 3D modeling has visualized the test outcome to know the characteristics that are needed specifically for the employee's turnover [6].

(Oliver et al. 2018) has provided a comparative study between supervised and semi supervised approach. According to their study supervised learning method are much better in performance as compared to semi-supervised learning methods [7].

(Lian, H., and Fan, Z., 2018) used Divide and Conquer approach to estimate L1-Penalized Support Vector Machine for classification. The statistical proof and inverse Hessian matrix minimize the gap between the theory and simulation used in this research study [8].

(vanEngelen, J.E., 2020) explored the survey study on supervised and semi supervised approaches. They have determined the various metrics used in evaluating the performances of these two approaches [9].

(Mondal K.C. et al. 2020) worked on three diverse data sets using automated ETL process. The ML approach reduce delay occurring between the data warehouse and Business Intelligence reporting tool [10].

3. Research Methodology

Supervised learning is perhaps the most efficient method in binary classification, since the objective is often to make the computer model of the classification algorithm. The objectives of this research work is -

- Analyze and implement scalability techniques with supervised algorithms.
- Predict the scalability of data with supervised algorithms.
- Analyze the scalability of data with performance, accuracy, F1 score and recall of the dataset with supervised machine learning algorithms.
- The confusion matrix determines the relation between the actual value and the predicted value for fitting the model under consideration based on training and test data sets.

For this study we have used a massive data set of Brain tumor. Before using the data for the training in the Machine Learning Algorithm and evaluating the test data, it is used the following steps to preprocess the data into usable format:

(1) Remove NULL values

(2) Columns with non-unique values excluded

(3) Removed unnecessary columns

(4) The outdated UTF format was modified to support UTF format

(5) The columns were rearranged according to weight of each column

(6) Save the cleaned data in csv file format

We have used python and machine learning programming language for implementation. We have used Anaconda software. Python is an effective programming language for the creation of a deep recurrent model of neural network. It supports in addition, a version of Python 3.7.1 used for this study is used. In the code creation process, Anaconda jupyter notebook python IDE is used; it is intended for operations in data analysis. A variety of science libraries, including Pandas, Numpy, Scipy, Matplotlib, sklearn and more, are available at Jupyter Notebook. It also provides advanced analysis, debugging, and editing capabilities in many apps, ("Jupyter Notebook: Anaconda Cloud”). The following libraries have been used such as numpy, pandas, matplotlib, scipy, seaborn and tensorf low etc. Malignant and Benign are the variable names used in the formation of confusion matrix.

The details of the two algorithms are given below-

3.1 Supervised Learning L1-Penalized SVM

Linear support vector machine (SVM) is the most specific binary identifications of machine learning methods. In compliance with applications in modern high-size statistics, we take account of problems with penalized SVM that involve minimizing the hinge loss function with a sparse convex regulator, such as, L1 on the variables, its categorized prediction and the sorted L1 penalty [11]. The current structure of algorithms is very nascent when composed to the normal linear L1-specific SVM — When the number of parameters and/or observations is high, each issue is represented as the Linear Program (LP) and is computer challenging. To that end, we are proposing new calculation algorithms for these LPs, putting together the

(1) conventional column generation techniques (and constraint) and

(2) first order non-linear convex methods for optimization which are rarely used in wide scale LPs together.

The comparable advantages of these elements are found to be effective as separate entities, but they are not simultaneously used to solve large-scale LPs [8].

3.2 Semi-Supervised Learning L1-Penalized SVM

Semi-supervised SVM are based on the development of the concept of margin maximizing to both identified and unidentifiable scenarios. In comparison to SVMs their formulation contributes to a problem of non-convex simulation. Recently a number of algorithms for solving S 3VMs were proposed. This paper addresses essential concepts in this literature. The behavior and performance of various S 3VM algorithms is intended to use the broad set of unlabelled data along with some labeled data to increase the performance of generalization. Naturally, a vigorously active topic was the creation of Support Vector Machines, which are able to handle partially-labeled data sets. One key task is to solve the regular SVM problem while considering unknown labels as external optimization variables. By optimizing the margins of unlabelled data, one learns a decision limit which crosses regions of small data size while respecting labels in the input spaces studied jointly in a joint experimental context [12].

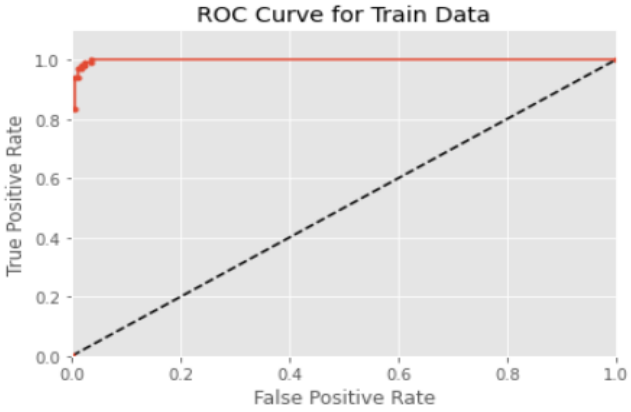
4. Data Analytics and Results

The various metric values for the training set and the test set under two algorithms are as follows:

4.1 ROC curve and Train Data Accuracy of L1-SVM

Figure 1 analyze that it is able it predicts overfit dataset and L1-SVM is a supervised algorithm. If n, we use relation sampling to limit the number of functions. In this reduced number of attributes, we implement the first order methods. We usually pick the top ten columns with the maximum absolute internal performance to be used for L1-SVM and Slope-SVM. For the Group-SVM issue: we measure the internal products for each group, between each function or within group and the answer and we take their L1-standard.

```
TRAIN DATA EVALUATION
---------------------------------------------
Average Train Accuracy: 0.986
Average Train Precision: 0.985
Average Train Recall: 0.993
Average Train F-Score: 0.988
Average Train AUC: 0.998
```



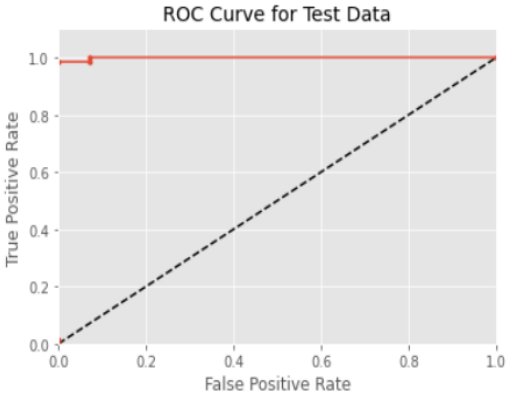

```
Confusion Matrix for Train Data:
+==================+==================+==================
|                  | Predicted Benign | Predicted Maligna
+------------------+------------------+------------------
|  Actual Benign   |       283        |          2
+------------------+------------------+------------------
| Actual Malignant |        5         |         164
+------------------+------------------+------------------
```

Figure 1.ROC curve and Train Data Accuracy of L1-SVM

Also, this figure analyzes that decision tree is predicting higher accuracy with 99% in comparison of random forest, less F1 score of 99%. From the output, it is clear that supervised algorithm works better with bigger datasets and overfit dataset.

4.2 ROC curve and Test Data Accuracy of L1-SVM

```
TEST DATA EVALUATION
-------------------------------------------
Average Test Accuracy: 0.964
Average Test Precision: 0.969
Average Test Recall: 0.974
Average Test F-Score: 0.971
Average Test AUC: 0.995
```




```
Confusion Matrix for Test Data:
+==================+==================+======================+
|                  | Predicted Benign | Predicted Malignant |
+------------------+------------------+----------------------+
|  Actual Benign   |        71        |          1           |
+------------------+------------------+----------------------+
| Actual Malignant |        1         |          42          |
+------------------+------------------+----------------------+
```

Figure 2. ROC curve and Test Data Accuracy of L1-SVM

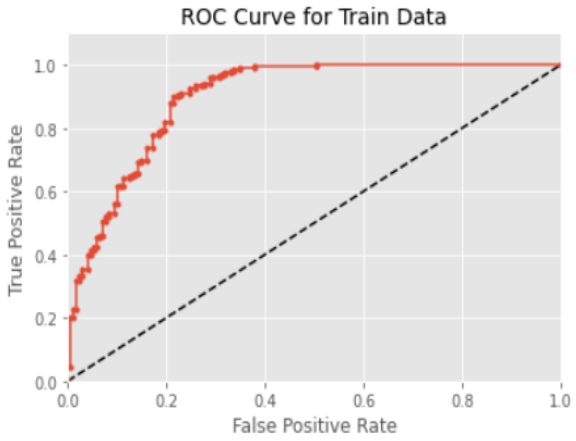
Figure 2, analyze thatL1-SVM issue with 20 regulator variables: We are compared to and without a warm-start with Gourbi's LP solver and to the tolerances of the column generation process. The simple LP solver is benefiting from the use of pleasant-start over β values — it requires many as 2 hours to conclude with P = 105 through warm-starts — CLG is now leading to over 2 500 times better results. CLG overall leads to major changes. The ROC curve predict that it is more equivalent and predicting higher accuracy. The LP on the full problem reaches the best objective values. The F1 scores of 97 percent and 100 percent are the clearest better accuracy. This is a misclassifying of more efficient classifiers, the L1 SVM algorithm worked very well with 100% greater accuracy.

The confusion matrix in both the cases provides the good fit of the model. This can be noticed by the diagonal values of the two variable under consideration.

4.3 ROC curve and Train Data Accuracy of semi L1-SVM

Figure 3, analyze that semi-L1-standard SVM has been used as initial value for such a recently introduced algorithm in solving the penalized non-convex SVM, and then it will be guaranteed in two iterative steps that an estimator possessing ultra-high dimensional oracle properties, implied in particular that the zero coefficients will be calculated as null at approximation of probabilities. Simulation studies show that the L1-standard SVM is a sparsely classified commodity and its efficiency can be used to resolve nonlinear SVM problems fined in the highest scale.

```
TRAIN DATA EVALUATION
--------------------------------------------
Average Train Accuracy: 0.852
Average Train Precision: 0.811
Average Train Recall: 0.997
Average Train F-Score: 0.894
Average Train AUC: 0.908
```




```
Confusion Matrix for Train Data:
+==================+==================+======================+
|                  | Predicted Benign | Predicted Malignant  |
+------------------+------------------+----------------------+
|  Actual Benign   |       284        |          1           |
+------------------+------------------+----------------------+
| Actual Malignant |        72        |          97          |
+------------------+------------------+----------------------+
```

Figure 3. ROC curve and Train Data Accuracy of semi L1-SVM

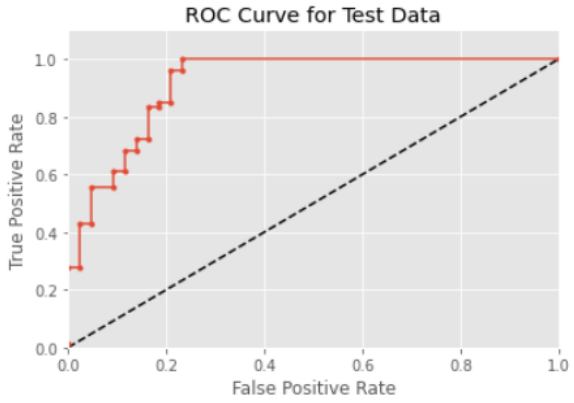
This figure also analyzes that the accuracy of semi L1-norm SVM algorithm. It is predicting 89% and is less accuracy with 100% in comparison with other algorithm. Semi L1-norm SVM is a regularized regression method often performing comparably higher with unsupervised algorithms and it has extra compelling advantage of being very effective. Semi L1-norm SVM introduces correlations that contribute to the effect of individual features on the likelihood of the occurrence of particular features on supervised algorithms. Thus, we are finally using semi L1-norm SVM in particular for multi-class classification to help us better understand the actual features to analyze and predict higher accuracy and results with our data set.

4.4 ROC curve and Test Data Accuracy of Semi L1-SVM

Figure 4, analyze the testing by semi L1-norm SVM algorithm. It initiates that an analysis of a baseline model on the dataset with L1 supervised algorithm. The baseline model includes 6 different variables. However, that several L1 regularization coefficients are neutralized, that are implemented in a semi L1-norm SVM algorithm to avoid over fitting model. This Figure also analyze score to optimal model analysis makes 100% reliable forecasts. The accuracy is not much higher in comparison of random forest algorithm. In the context of the dataset, however, it is not appropriate for overfit data. The algorithm of supervised algorithm is that it can predict the presumption of independent variables. Semi L1-norm SVM algorithm believes that all of the features are statistically independent and it can predict overfit data.

The confusion matrix here also in both the cases provides the good fit of the model.

```
TEST DATA EVALUATION
--------------------------------------------
Average Test Accuracy: 0.858
Average Test Precision: 0.817
Average Test Recall: 0.997
Average Test F-Score: 0.898
Average Test AUC: 0.914
```




```
Confusion Matrix for Test Data:
+==================+==================+=================
|                  | Predicted Benign | Predicted Malign
+------------------+------------------+-----------------
|  Actual Benign   |        72        |          0
+------------------+------------------+-----------------
| Actual Malignant |        10        |         33
+------------------+------------------+-----------------
```

Figure 4. ROC curve and Test Data Accuracy of semi L1-SVM

5. Conclusion

The scope of this research work renders around the massive data set of tumors. Supervised algorithms require a detailed fine tuning of the variables and a considerable number of situations for the data collection. It constructs the algorithm model only with precision and accurate classification. We have taken two novel algorithms namely supervised L1 SVM and semi L1 SVM. The confusion matrix in all cases indicates the a good fit of the model under consideration. From the results, it is concluded that L1-norm SVM algorithm able to work well with scalable and large dataset in comparison of other algorithm. Thus this research work highlight that the L1 SVM supervised algorithm works well with scalable data whereas Semi L1 SVM supervised algorithm works well with over fit data. Thus study provided by (Oliver et al. 2018) [7], has been well analyzed on the modified formats of the two mentioned approaches. The future scope of this research work can be extended to the ELT tools for the mentioned metrics of the machine learning algorithms.

References

1. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., 2017. Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), pp.128-138.
2. Markus G., Matthias R., Christian B., Gabriele C., Philipp G., Morris R. and J´on Atli B. 2015. On Scalable Data Mining Techniques for Earth Science, Elsevier- Procedia Computer Science, volume 5, 2188–2197
3. Gupta, Preeti, Arun Sharma, and Rajni Jindal, 2016. "Scalable machine learning algorithms for big data analytics: a comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6, no. 6: 194-214.

4. F. Faghri and Sayed Hadi Hashemi and M. Babaeizadeh and Mike A. Nalls and Saurabh Sinha and R. Campbell, 2017. Towards Scalable Machine Learning and Data Mining: The Bioinformatics Case, ArXiv- Computer Science, Mathematics, Volume 1710.00112, 1-9.
5. Quan Zou, Dariusz Mrozek, Qin Ma, and Yungang Xu, 2017, Scalable Data Mining Algorithms in Computational Biology and Biomedicine, Hindawi BioMed Research International Volume 2017, Article ID 5652041, 1-3 pages https://doi.org/10.1155/2017/5652041
6. Alam, M. M., Mohiuddin, K., Islam, M. K., Hassan, M., Hoque, M. A.-U., & Allayear, S. M. 2018. A Machine Learning Approach to Analyze and Reduce Features to a Significant Number for Employee's Turn Over Prediction Model. Intelligent Computing, 142–159. doi:10.1007/978-3-030-01177-2_11
7. Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., Goodfellow, I. J. (2018). Realistic evaluation of deep semi-supervised learning algorithms. arXiv:1804.09170.
8. Lian, H, and Fan, Z., 2018, Divide-and-Conquer for Debiased l1-norm Support Vector Machine in Ultra-high Dimensions, Journal of Machine Learning Research, 18, pp. 1-26.
9. van Engelen, J.E., Hoos, H.H. (2020) A survey on semi-supervised learning. Mach Learn 109, 373–440. https://doi.org/10.1007/s10994-019-05855-6
10. Mondal K. C., Biswas, N. and Saha, S., 2020. Role of Machine Learning in ETL Automation, ICDCN 2020: Proceedings of the 21st International Conference on Distributed Computing and Networking, Article No.: 57, pp. 1–6, https://doi.org/10.1145/3369740.3372778
11. Crisci, C., Ghattas, B. and Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling, 240, pp.113-122.
12. Jain, P., Garibaldi, J.M. and Hirst, J.D., 2009. Supervised machine learning algorithms for protein structure classification. Computational biology and chemistry, 33(3), pp.216-223.
13. Brusilovsky, P. Chavan, G., &Farzan, R. (2004). Social adaptive navigation support for open corpus electronic textbooks. Adaptive Hypermedia and Adaptive Web-Based Systems,3137, 24–33.
14. Cottrell, D. M., & Robinson, R. A. (2003). Blending learning in an accounting course, *The Quarterly Review of Distance Education*, *4*(3), 261–269.
15. Graham, C. R. (2005). *Blended learning systems: definition, current trends, and future directions, The handbook of blended learning global perspectives.* San Francisco: Pfeiffer Publishing.