Ensemble Framework for Big Data Stream Mining

Phyo Thu Thu Khine
University of Computer Studies, Hpa-an
Hpa-an, Myanmar
phyothuthukhine@gmail.com

Htwe Pa Pa Win
University of Computer Studies, Hpa-an
Hpa-an, Myanmar
hppwucsy@gmail.com

Abstract

The rapid development of industry enterprises, the large amount of data generated by these originalities and the exponential growth of industrial business website are the causes that lead to different types of big data and data stream problem. There are many stream data mining algorithms for classification and clustering with their specific properties and significance key features. Ensemble classifiers help to improve the best predictive performance results among these up-to-date algorithms. In ensemble methods, different kinds of classifiers and clusters are trained rather than training single classifier. Their prediction machine learning results are combined to a voting schedule. This paper presented a framework for stream data mining by taking the benefits of assembling technology based on miss classification stream data. Experiments are carried out with real world data streams. The experimental performance results are compared with the modern popular ensemble techniques such as Boosting and Bagging. The increasing in accuracy rate and the reducing in classification time can be seen from the test results.

Keywords: *Big Data, Bagging, Boosting, Data Stream Mining, Ensemble Classifiers, Misclassification Stream Data*

I. INTRODUCTION

In the strong-growing of big data era, all the internet application significantly needs to process large amount and varieties of data. This growing is quickly rapid up and affecting to all technology and businesses environments for organizations and individuals benefits respectively. Furthermore, big data analysis intend to extract the statistical information using data mining algorithms in instantaneously way that assist in making likelihoods, finding the hidden information, classifying recent developments and defining decisions. Though, the rise in classification speed comes at what cost, difference in estimation with the original, and mis-assigning in relative classes whatever machine learning algorithms used [1-3].

To overcome this problem, this paper focuses on finding a way to speed up the mining of streaming in high accuracy rate based on miss classification data streams.

Section 2 deals with the nature of big data and how it is associated in real-world applications. Section 3 describes about the data stream mining, and briefly describes modern data stream mining methods. Section 4 provides the previous research for data stream classification. Section 5 proposes the classification framework. Experimental setups and results are depicted in Section 6. Finally, conclusions are presented in section 7 to summarize outcomes.

II. BIG DATA

The meaning of "Big Data" can be classified into many ways: someone defines that big data is the large amount of data over a certain threshold. Others defined as data that cannot handle by the conventional analytical suits such as Microsoft Excel. More popular mechanisms identified big data as data that has the Variety, Velocity, and Volume features. Big data analytics is an innovative approach including of various mechanisms and procedures to extract treasured insights from raw data that does not suitable for the traditional database system due to any reasons.

Big data applications can be found in several fields such as financial area, technological area, electronic governmental area, business and health care processes, etc. Furthermore, in other specific cases, energy control used big data, anomaly detection, crime prediction, and risk management. Big data are having a strong control for every kinds of business.

Information data can be defined as a new form of investment, a different type of currency, and an original resource of valuable things. It has been revealed about the power of big data that has the efficient strategies to become successful business. But it can't be argued that all the strategies of big data may not be used for all business types. However, it is the universal truth that a data information strategy is still valuable, whatever the size of data. This enormous

amount of data in application opens new challenging detection tasks and lead to Data Stream Mining [4].

III. DATA STREAM MINING

In computer science, data stream mining is associated with two fields: data mining and data streams. It turns out to be essential areas of computer science applications such as industrial engineering processes, transaction flows of credit card, robotics, e-commerce business, spam filtering, sensor networks, etc.

The data stream mining task quite different traditional data mining task regarding processing or executing the mining task, but the objectives are the same. The normal algorithms of data mining methods cannot be used directly for data streams because of the following factors:

-Data Streams may be large amount of data and these are actually unlimited number of elements.

-Data Streams can be arrived to the system in a short period of time.

-Data Streams may be changed to different manners during the distribution times of processing.

Therefore, the algorithms for data stream need to store previous information in a condensed format structures. The most widespread methods for the data stream classification are categorized into the following groups;

- Instance-Based Learning Methods
- Bayesian Learning Methods
- Artificial Neural Networks Learning Methods
- Decision Trees Learning Methods
- Ensemble Learning Methods
- Clustering Methods

A. Instance-Based Learning Methods

Instance-based learning classifiers are also called k-nearest neighbors learners. As these instance classifiers can process incremental learning method, there is a necessary to store all the previous data elements in the memory. Therefore, these normal learning methods can't be directly used for data streams [5]. All the series of instance-based classification algorithms were presented in [7].

B. Bayesian Learning Methods

Bayesian learning classification methods are based on the standard Bayesian theorem. The aim of Bayesian learning is to evaluate the essential likelihoods using the existing training dataset. Then, a learning algorithm is used to categorize new data—the group which maximizes the next probability is allocated to an uncategorized or unlabeled element. Naive Bayes learning method is done in an incremental fashion. However, they need to have a fix size of memory. These Naive Bayes learning features possibly appropriate in the mining process of data stream [5].

C. Artificial Neural Networks Learning Methods

Artificial Neural Networks learning methods are likely the nervous system of the animals. Multi-Layer Perceptron is the most common learning classification method. When the number of data streams training elements is large, neural network learning can be transformed to single-pass incremental way. If input neurons and synapses number is kept unaltered during the learning process then the memory requirement is kept constant. The above properties of neural networks can be appropriate for data streams [5, 8].

D. Decision Trees Learning Methods

The state-of-the-art Decision Tree algorithms can be used for classification of data stream. The algorithms in this type are based on Hoeffding trees method. For the static data, Hoeffding tree chooses an attribute that is appropriate to split tree nodes. Because of the infinite size of the data streams, all the data elements of the node can't be kept in memory. Therefore, evolutionary learning algorithms are used for data streams. The most noticeable method of this type of learning is the VFDT algorithm [10].

E. Ensemble Learning Methods

Other learning techniques which can be applied for data stream mining are ensemble ways. Various approaches are suggested to combine many single algorithms into a group to form ensemble classifiers. Among the state-of-the-art classifiers available in many data mining environment, including the stream data mining, assembling of classifiers provides the best performance [5, 11-13]. The most common and effective approaches of ensemble methodologies are Bagging and Boosting. These popular methods have been discovered in the data stream scenario and are firstly presented by Oza and Russell [14, 15].

F. Clustering Methods

Clustering can be used for the unsigned instances that have homogeneous clusters relating with

their similarities. Streaming methods for clustering can be done with two levels, online and to eliminate similar data which in offline. At the online level, a set of extremely small clusters is computed and updated from the stream efficiently; in the offline phase, a classical batch clustering method, for example, k-means is performed on the micro clusters. Although the offline level clustering carries out for several amount of processing phase, the online level clustering only done with a single pass phase for the input data. Because the offline processing can be separated to a set of small clusters and can be invoked when the stream ends. Furthermore, they can update the group of separated clusters periodically according to the stream flows they need.

The k -means clustering method is one of the most used methods in clustering, due to its simplicity. To initiate the clustering process, the value of k is chosen in a random way, but most popular developed algorithms begins with 1, and some starts with 5 or 10. Then, according to that centroid value, each instance is assigned to the nearest centroid. The cluster centroids are computed again with the center of mass of the assigned instance. This process is computed repeatedly until the desired criterion is encountered or the assignments cannot be changed. This routine cannot be used for data streams mining process because the streams require many passes to be clustered [16].

However, this paper focuses on Bayesian Classifiers and Ensemble Classifiers and Decision Trees and Cluster.

IV. Literature Review

Hundreds of academic papers have been presented based on research done for big data classification on the standard dataset of data stream mining. These works done are described according to categorization of the above state-of-the-art classification groups.

The author in paper [6] illustrates the developed numerous streaming data computation platforms and discuss their major abilities. They clearly specify the prospective research directions for high-speed large-scale mining methods for data stream from different point of views such as procedures, implementation nature and performance evaluation analysis. They clearly described that Instance-Based Classifiers get more accuracy among the other classifiers but the time taken for that is extremely large. Therefore, in order to perform faster changes, the authors in [17] took the distributed computing advantages and then proposed the nearest neighbor incremental classifier.

In [18], an operational pattern-based Bayesian learning classifier was proposed to handle data streams. The researchers in [19] implemented highly efficient and popular algorithm “Naïve Bayes Algorithm” on huge complex data to acquire knowledge. They proposed reduction technique to remove similar data which in sequence reduces the time of computation, the amount of memory space requirement, and enhances the performance of Naïve Bayes Algorithm. Their research work indicates highly efficient Naive Bayes Algorithm solution or huge data streams.

Many authors proposed to apply Neural Network Deep Learning methods for data stream processing in many ways. Neural Network deep learning architectures can be capable for complex tasks and sometimes it can outperform human’s beings in some application areas. Although the remarkable advancements for this area can be seen clearly, there is an ill-posed optimization problem for training deep architectures that has a very large number of hyper-parameters. For this reason, modification of Neural Network integrated framework has been presented to get online calculation capabilities of highly scalable solution for mining data streams in [20].

Extensive surveys research on assembling for classification of data stream and also regression tasks have been done by [13]. They surveyed wide ranges of data streams ensemble techniques and introduces the innovated learning methods for imbalance data streams processing including complex representation of data, semi-supervised learning, the structured outputs and the detection method. The authors in [16] propose a new ensemble learning method, called Iterative Boosting Streaming ensemble (IBS) that can be able to classify streaming data. The authors in [21] introduce the new distributed training model for ensemble classifiers to avoid the dangers of the vote-based separated ensembles. This model is named as “LADEL”.

V. Proposed Framework

Let a data stream input as a sequence of batches DS = {S1, S2 . . ., St,}, where St = {S1…, SN} be an unlabeled batch. Assume the real class label Li of instance Si, for i = 1. . . N, and the equivalent labeled set, defined as ˆSt = {(S1, L1) . . ., (SN, LN)}, that can be used at the training stages. The class labels are necessary to be predicted manually for incoming unlabeled data of real-world data. The standard stream dataset no need to do this state. An automatic mining system of data streams, that has acceptable and

constant performance at classification accuracy, computation procedure and memory usage.

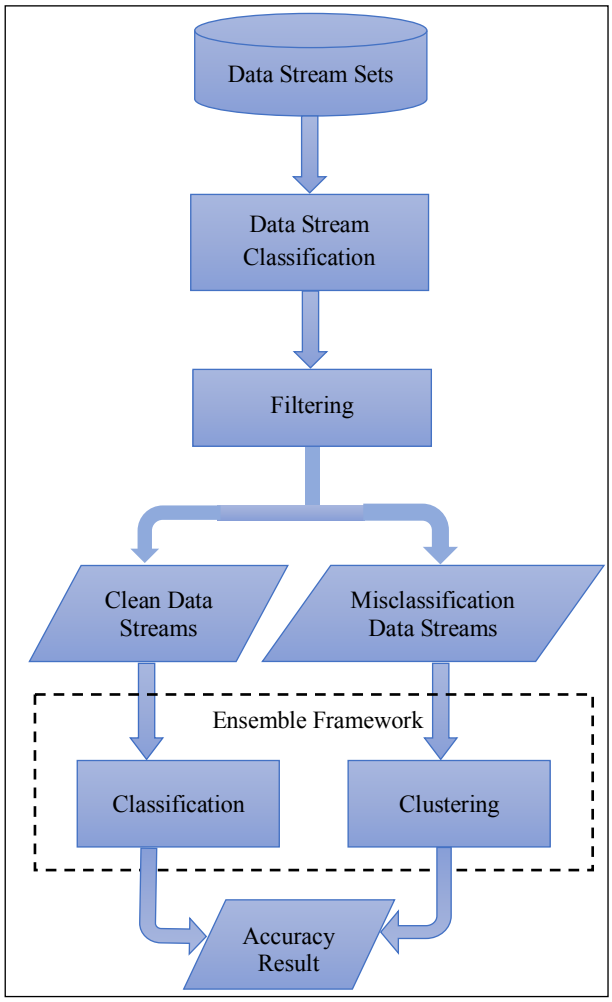

Figure 1. Ensemble Framework for Data Streams

After labeling batches, St, is presented, the simple classifier categorizes its instances. There is an assumption that the name of the labels will be known immediately after the classification process took place, so the miss classification error of this batch can be predicted and clearly know the correct instance stream. The uncertain, ambiguous, incomplete, and subjective data can reduce the performance of the classifier and not all the techniques are suitable for all data streams. Therefore, the misclassification data streams are filtering out after the classification process. After the filtering process is created, assembling method is designed by using the simple cluster to improve the model to focus on data stream that is not easy to classify. The miss classified data streams, MDS = {S1, S2… St} are separated from correctly labeled streams, clean data streams CDS= {S1, S2…, St} from DS. Then the clustering is ensemble to label the left incorrect data streams, MDS. After grouping batches of MDS, the accuracy for that clusters is calculated, and the overall accuracy of the DS can also be calculated as illustrated in Fig. 1.

VI. EXPERIMENT RESULTS

The well-known data set of streams, real world Electricity [22] data is used to test the ensemble effect. The Electricity data was accumulated from the New South Wales' electricity market, Australian State. Prices are unstable and depends on the demand and supply in this market. It contains the real data collected at every 30 minutes for 2 years and 7 months. This dataset consists of 45,312 instances with five attributes for the time, day, period and price. The class label defines the alteration of the price corresponding to moving average (MA) of the past 24 hours.

The experiments are conducted on the tasks of data stream classification. The experiment design is implemented with the use of EvaluatePrequential approach because the system needs to know the label of the class and to filter out the mis-classified data. Firstly, the classification is done with the well-known classifier Naïve Bayes and VFDT and the results are shown in Table I.

TABLE I. PERFORMANCE MEASUREMENT COMPARISON FOR SINGLE STANDARD DATA STREAM CLASSIFIERS

Name of Data Stream Algorithm	Classification Accuracy (%)	Kappa Statistic	Kappa Temporal Statistic	Elapsed Time (s)
Naive Bayes	73.07	40.89	-83.57	0.67
VFDT	72.23	43.59	-89.30	0.52

The data streams are tested with the standard ensemble methods of Leveraging and Boosting. The results from the experiments are summarized in Table II.

TABLE II. PERFORMANCE MEASUREMENT COMPARISON FOR ENSEMBLE DATA STREAM CLASSIFIERS

Name of Data Stream Algorithms	Classification Accuracy (%)	Kappa Statistic	Kappa Temporal Statistic	Elapsed Time (s)
LeveragingNB	52.82	12.95	-221.62	1.58
LeveragingVFDT	75.497	48.38	-67.04	4.25
OZOBoostNB	74.322	44.38	-75.04	1.02
OZOBOOSTVFDT	69.352	39.70	-108.92	1.72

Then experiments are carried out for the part of the proposed framework by using the simple KMeans clustering method. The data and results are shown in Table III and this addition tasks can enforce the

classification accuracy but need to care for elapsed time. The overall performance for the proposed ensemble method is illustrated in Table IV.

TABLE III. PERFORMANCE MEASUREMENT COMPARISON FOR CLUSTERING FOR MISS DATA

Data Stream	Classification Accuracy (%)	Data Count	Elapsed Time (s)
Miss data for NB	45.57	12202	0.03
Miss data for VFDT	59.29	12583	0.08

TABLE IV. PERFORMANCE MEASUREMENT COMPARISON FOR PROPOSED ENSEMBLE DATA STREAM CLASSIFIERS

Name of Data Stream Algorithms	Classification Accuracy (%)	Elapsed Time (s)	Improved Accuracy (%)
NB + KMeans	88.04	0.7	14.97
VFDT + KMeans	88.69	0.6	16.46

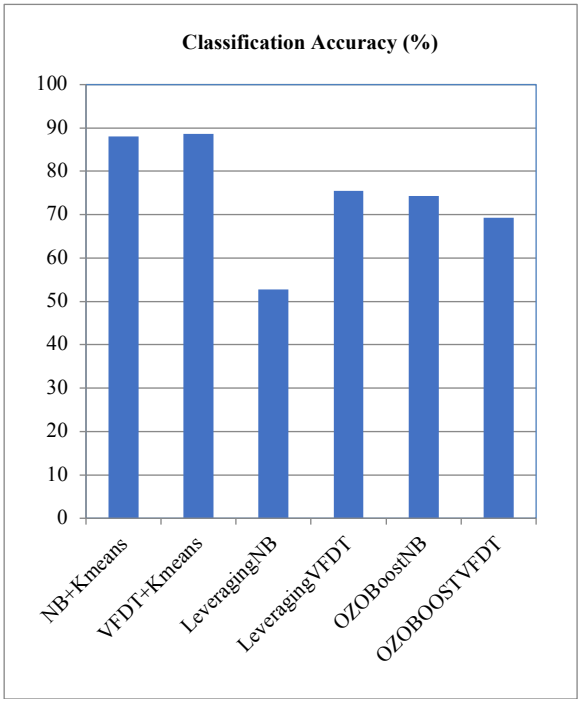

Figure 2. Measurement Comparison for proposed Ensemble Data Stream Classifiers and Standard Ensemble Classifier

Then the comparison is carried out the proposed ensemble method and the state-of-the-art ensemble methods and results are shown in Fig. 2 and Fig. 3. From these results, it can be seen clearly that the proposed framework not only can increase the classification accuracy but also less than in elapsed time.

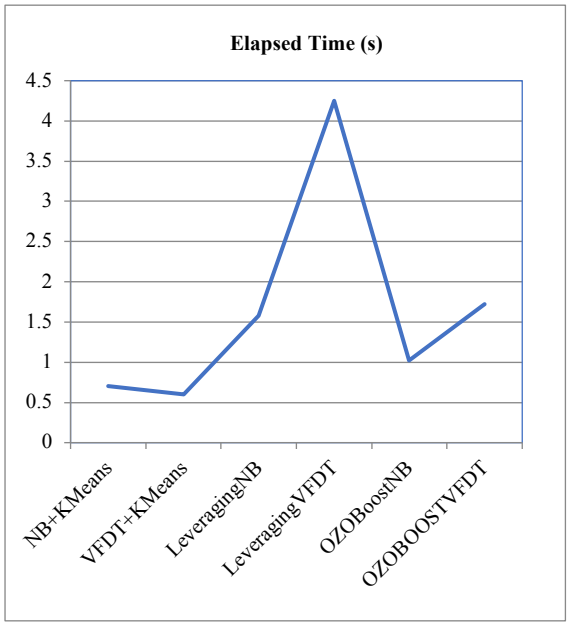

Figure 3. Time Consume comparison for proposed Ensemble Data Stream Classifiers and Standard Ensemble Classifier

VII. CONCLUSION

In this paper, the ensemble framework constructed from the data streams classifiers and simple K-Means clustering is proposed for mining data streams. The proposed framework of the ensemble learning classifiers, the combination of Naïve Bayes and K-Means, and VFDT and K-Means, has been evaluated. Furthermore, the comparison of the proposed framework against state-of-the-art ensembles, Leveraging and Boosting using standard data stream set. The results clearly show that the proposed framework not only can improve the classification accuracy based on mis-classification data, but also can reduce the time taken than the above standard ensemble techniques. Future research will concentrate on learning the influence of the size of stream data and more effective ensemble mechanisms on accuracy of the ensemble classifier.

REFERENCES

[1] N. Sun, B. Sun, J. Lin and M. Yu-Chi Wu, "Lossless Pruned Naive Bayes for Big Data Classifications," Big Data Research, vol.14, pp. 27-36, December 2018. https://doi.org/10.1016/j.bdr.2018.05.007.

[2] C. Tsai, C. Lai, H. Chao and A. V. Vasilakos, "Big Data Analytics: A Survey," Journal of Big Data, vol.2, A21, pp. 1-32, December 2015. https://doi.org/10.1186/s40537-015-0030-3.

[3] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, A., I. Abaker Targio Hashem, A. Siddiqa, and I. Yaqoob, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," IEEE Access, vol.5, pp. 5247-5261, May 2017. https://doi.org/10.1109/ACCESS.2017.2689040

[4] F. Corea, An Introduction to Data Everything You Need to Know About AI, Big Data and Data Science. ISBN 978-3-030-04467-1, Springer Nature Switzerland AG, 2019. https://doi.org/10.1007/978-3-030-04468-8.

[5] L. Rutkowski, M. Jaworski and P. Duda, Stream Data Mining: Algorithms and Their Probabilistic Properties, Studies in Big Data, Volume 56, ISSN 2197-6503, Springer Nature Switzerland AG: Springer International Publishing, 2020.

[6] B. Rohit Prasad and S. Agarwal, "Stream Data Mining: Platforms, Algorithms, Performance Evaluators and Research Trends," International Journal of Database Theory and Application, vol. 9, No. 9, pp. 201-218, 2016. http://dx.doi.org/10.14257/ijdta.2016.9.9.19.

[7] D.W. Aha, D. Kibler and M.K. Albert, "Instance-Based Learning Algorithms," Machine Learning, vol. 6, No. 1, pp. 37-66, 1991. https://doi.org/10.1007/BF00153759.

[8] J. Gama, P. Pereira Rodrigues, "Stream-Based Electricity Load Forecast," Knowledge Discovery in Databases: PKDD 2007, Lecture Notes in Computer Science, vol. 4702, pp. 446-453, 2007, Springer, Berlin. https://doi.org/10.1007/978-3-540-74976-9_45.

[9] D. Jankowski, K. Jackowski and B. Cyganek, "Learning Decision Trees from Data Streams with Concept Drift," International Conference on Computational Science 2016, ICCS 2016, San Diego, California, USA, 6–8 June 2016. Procedia Computer Science vol. 80, pp. 1682-1691, 2016. https://doi.org/10.1016/j.procs.2016.05.508.

[10] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, USA, pp. 71-80, 2000. https://doi.org/10.1145/347090.347107.

[11] J. N. van Rijn, G. Holmes, B. Pfahringer and J. Vanschoren, "The Online Performance Estimation Framework: Heterogeneous Ensemble Learning for Data Streams," Machine Learning, vol. 107, No. 1, pp. 149-176, 2018. https://doi.org/10.1007/s10994-017-5686-9.

[12] L. I. Kuncheva, "Classifier Ensembles for Detecting Concept Change in Streaming Data: Overview and Perspectives," In Proceedings of the 2nd Workshop SUEMA, ECAI, Patras, Greece, pp. 5-9, July 2008.

[13] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski and M. Wo´zniak, "Ensemble Learning for Data Stream Analysis: A Survey," Information Fusion, vol. 37, pp. 132-156, 2017. https://doi.org/10.1016/j.inffus.2017.02.004.

[14] N. C. Oza and R. Russell, "Online Bagging and Boosting," In Eighth International Workshop on Artificial Intelligence and Statistics, pp. 105-112, January 2001, Morgan Kaufmann, Key West, Florida, USA.

[15] J. Roberto Bertini Junior and M. Carmo Nicoletti, "An Iterative Boosting-Based Ensemble for Streaming Data Classification," Information Fusion, vol. 45, pp. 66-78, 2018. https://doi.org/10.1016/j.inffus.2018.01.003.

[16] A. Bifet, R. Gavaldà, G. Holmes and B. Pfahringer, Machine Learning for Data Streams: with Practical Examples in MOA. ISBN: 9780262037792. The MIT Press, 2018.

[17] S. Ram´ırez-Gallego, B. Krawczyk, S. Garc´ıa, M. Wo´zniak, J. Manuel Ben´ıtez and F. Herrera, "Nearest Neighbor Classification for High-Speed Big Data Streams Using Spark," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, issue. 10, pp. 2727-2739, 2017. https://doi.org/10.1109/TSMC.2017.2700889.

[18] J. Yuan, Z. Wang, Y. Sun, W. Zhang, and J. Jiang, "An Effective Pattern-based Bayesian Classifier for Evolving Data Stream," Neurocomputing, pp. 1-12, 2018. https://doi.org/10.1016/j.neucom.2018.01.016.

[19] S. David, K. Ranjithkumar, S. Rao, S. Baradwaj, and D. Sudhakar, "Classification of Massive Data Streams Using Naïve Bayes," IAETSD Journal for Advanced Research in Applied Sciences, vol. 5, issue 4, pp. 208-215, 2018.

[20] M. Pratama, P. Angelov, J. Lu, E. Lughofer, M. Seera and C. P. Lim, "A Randomized Neural Network for Data Streams," International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, May 2017, pp. 14-19. https://doi.org/10.1109/IJCNN.2017.7966286.

[21] S. Khalifa, P. Martin, and R. Young, "Label-Aware Distributed Ensemble Learning: A Simplified Distributed Classifier Training Model for Big Data," Big Data Research, vol. 15, pp. 1-11,2019. https://doi.org/10.1016/j.bdr.2018.11.001.

[22] M. Harries, "Splice-2 Comparative Evaluation: Electricity Pricing," Technical Report 9905, School of Computer Science and Engineering, University of New South Wales, Australia, 1999.