Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
ISSN: 2788-7421

Big Data Classification Efficiency Based on Linear Discriminant Analysis

Ahmed Hussein Ali[1*], Zahraa Faiz Hussain[2], Shamis N. Abd[2]

[1]Department of computer, College of Education, AL-Iraqia University, Baghdad, IRAQ

[2]Computer Science Department, AL Salam University College, Baghdad, IRAQ

*Corresponding Author: Ahmed Hussein Ali

DOI: https://doi.org/10.52866/ijcsm.2019.01.01.001
Received October 2019; Accepted October 2019; Available online January 2020

ABSTRACT: The proliferation of online platforms recently has led to unprecedented increase in data generation; this has given rise to the concept of big data which characterizes data in terms of volume, velocity, variety, and veracity. One of the common multivariate statistical data analysis tools is linear discriminant analysis (LDA) which relies on the concept of obtaining the separation among groups through LDA. The prediction of the class of a given class of data points can be achieved through classification, a supervised learning technique but prior to a classification process, a classification model must first be built using classification algorithms. Several classification algorithms are available for prediction tasks. LDA is commonly used for the reduction of the dimensionality of datasets. In this article, the use of LDA to improve the classification performance of different classification model was presented.

Keywords: Spark, Parallel Processing,, Dimensionality Reduction.

1. INTRODUCTION

Internet users are more attracted recently to micro-blogging websites and online social networks (OSN)compared to any other kind of website. Hence, the rate of data generation on these platforms has continued to grow both in speed, volume, velocity, and variety, giving rise to the era of big data [1]. Numerous researchers have paid attention to big data [2] with the intent of establishing the pattern of people‘s opinions and the distribution of users in these digital platforms. It is believed that people easily use these websites to share their opinions on the things that interest them; however, the increasing number of comments and posts made on these sites has made it difficult to regulate their content. Hence, data mining has been introduced as a [3] tool for knowledge discovery from large data sets. Classification remains a major problem during data mining. Dimensionality reduction is a process of minimizing the total number of random variables in a manner that the principal variables are obtained. LDA [4] is commonly applied on linear datasets for dimensionality reduction. To predict class labels, first, the models or classifiers must be constructed, and during classification tasks, too many features are often seen, and the classification process is performed on these features. The presence of too many features increases the training complexity of the model; so, it is necessary to reduce the dimension of the features to improve the efficiency of the classification model. The performance of classification models is evaluated based on certain performance evaluation metrics, such as accuracy, ROC curve, specificity, and sensitivity. This work presents the development of four classification models based on three different datasets. The classification models in this study were built using four different classification algorithms which are Kernel Approximation Algorithm, J48, Stochastic Gradient

Descent Algorithm, and Multilayer Perceptron. The dimensionality of the employed datasets was reduced using LDA and the performance metrics were evaluated before and after LDA implementation.

2. RELATED WORK

Several big data processing systems have been proposed by the research community; this section focuses on the latest works in the field of classification and dimensionality reduction for big data. Shuiwang et al [5] presented a unified framework for generalized LDA via a transfer function. There are four steps of the proposed framework, and the study proved the differences in the transfer functions of various generalized LDA algorithms. The combination of different algorithms exploits the differences and commonalities of generalized LDA frameworks. ULDA is specifically known to be a combination of PCA, LDA, & RLDA. However, the overfitting problem of the ULDA has been analyzed to show how the different components (RLDA, PCA, and LDA) overcome this issue by applying PCA, regularization, and dimensionality reduction. Delin et al [6] presented a new batch LDA algorithm called LDA/QR that related closely with the ULDA [6]. The new model was developed by computation of the economic QR factorization of the data matrix, followed by solving a lower triangular linear system. The performance of classifiers before and after application of LDA has been investigated by Joyoshree et al [7]; the analysis was done as a technique for dimensionality reduction. The study relied on datasets sourced from the UCI ML repository; four different classifying algorithms were applied to each dataset. Jieping proposed a new formulation of the generalized LDA based on trace optimization [8]. The optimization problem was solved using generalized singular value decomposition while the solution from the LDA/GSVD is considered a special solution to the optimization problem. The theoretical equivalence between the Pseudo Fisher LDA (PFLDA) and the exact algorithm was demonstrated, thereby justifying the pseudo-inverse based LDA.

3. SPARK ECOSYSTEM

Owing to the rich expressive capability of RDD [9, 10], it has been the basis for four major models which are Spark SQL [11], Spark Streaming, MLlib, & GraphX [12]; these four models, together with the basic components of Spark, made up a simple Spark framework (see Figure 1). Each of the core parts is introduced as follows: Spark Core: its main trick is its rich operational interface and the RDD data structure; Spark SQL: it is the module for the implementation of Spark with structured data; Spark Steaming: Enables the building of a fault-tolerant and extensible streaming application. As streaming computing is broken into several short batch jobs, the lightweight and low-latency scheduling framework of Spark also supports streaming efficiently. Spark MLlib [13, 14]: it is an extensible ML library for Spark that contains the tools and common learning algorithms, such as linear regression, binary classification, gradient descent, low-level optimization primitives, and clustering, collaborative filtering. Spark GraphX: a new Spark API that supports parallel graph computing and graph computing on the Spark platform.

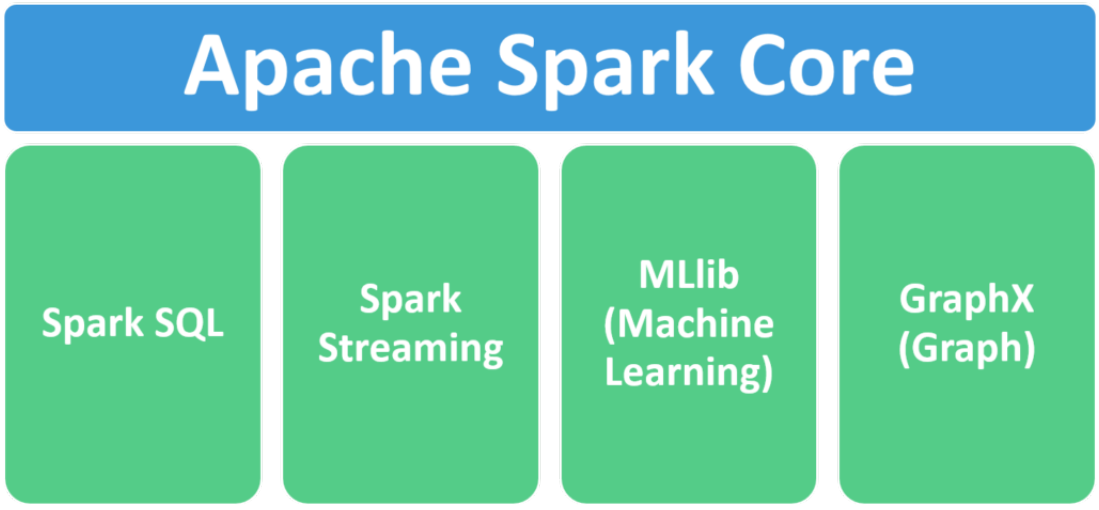

FIGURE 1. Spark overall structure

4. LINEAR DISCRIMINANT ANALYSIS (LDA)

The LDA [15, 16] is a common technique that has been used for years as a dimension reduction method for clustered data in pattern recognition; it is formulated as an optimization problem on scatter matrices. LDA is faced with the challenge of its objective function requiring that the total scatter matrix be nonsingular. LDA is also called Fisher Discriminant Analysis (FDA); it seeks how to achieve efficient discrimination, and this differentiates it from PCA which seeks how to achieve efficient representation. Both techniques can be implemented only when the complete training dataset has been provided in advance, while learning is performed on another batch. Both methods are widely used in mobile robotics and face recognition; they are also used in some knowledge discovery and data mining applications. When conducting LDA/PCA learning on datasets in the real world, some difficult situations are encountered, such as a situation where the training set is not provided beforehand. Being that data representation in most cases is done as a discriminant analysis (seeFigure 2), coupled with the fact that we may not be aware of the kind of data to be presented in the future, it means that both the individual samples and scale of the chunks in each chunk are randomly given. A common method for data processing is data dimensionality reduction but one of the effective methods is a linear transformation where the high dimensional raw data is projected into a low-dimensional space. LDA is a supervised method of dimensionality reduction that aims to project a larger inter-class distance and a smaller intra-class distance in the projected space.

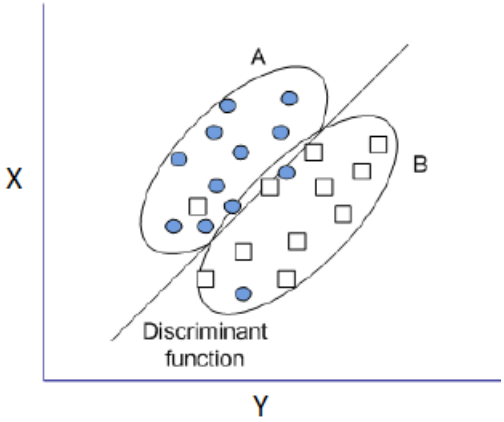

FIGURE 2. **Discriminant Function**

5. CLASSIFYING ALGORITHM AND PERFORMANCE EVALUATION

In data mining, classification is one of the ways of predicting, classifying, and organizing data into appropriate labels, with each label/class having its own rules & algorithms. Most of the existing classification models follow decision tree rules, such as J48, while some follow the Bayesian Network, such as Naïve Bayes and Bayes Net. Some have also been designed to follow artificial intelligence and NN. There are different applications of classification techniques, and each requires a specific class of dataset. Furthermore, the best classification techniques for a set of testing data can be determined by ensuring the compatibility of the data with the rules and algorithms of the selected technique since there is no general classification method that can perform efficiently on all datasets.

Stochastic Gradient Descent Algorithm [17] was developed as an iterative technique for the optimization of an objective function with appropriate smoothness characteristics (such as differentiable or sub-differentiable). It is considered a stochastic approximation of the gradient descent optimization because its process replaces the actual gradient that is estimated from the whole data set with an estimate that is estimated from a subset of the data that has been randomly selected. This reduces the computational burden in high-dimensional optimization problems, thereby making iteration faster in exchange for a lower convergence rate.

Kernel Approximation Algorithm [18]: This is an algorithm mainly designed for pattern analysis; the commonest member of the kernel machines is the support-vector machine (SVM). During pattern analysis, the major aim is to find and study the types of relationships existing in a given dataset, such as clusters, principal components, rankings, correlations, classifications, etc. Most of the algorithms used to solve this task require the explicit transformation of the data in raw representation into feature vector representations using a user-specified feature map. On the other hand, the kernel methods only require a kernel specified by the user, i.e., a similarity function over data point pairs in raw representation.

J48 [19] was developed as an optimized version of C4.5. It is based on a Decision tree and remains one of the common classification methods in data mining and data processing. The rules and algorithm of J48 rely on decision tree techniques that recognized the existence of leaves and branches, with each leaf or branch containing a decision that offers a different solution. Some datasets have large tree models with numerous leaves that offer different solutions compared to those with few leaves as found in the J48.

Multilayer perceptron [20] is based on NN and Artificial intelligence without qualification. An MLP has at least three layers, made up of the input, hidden, and output layers. It is a feedforward NN and may contain one or more hidden layers. Each MLP node in each layer is linked to all the nodes of the layer. MLP is among the data classification methods used in NN, DL, and other data processing applications.

In this study, the classification dataset was sourced from the UCI ML data repository and the features of the dataset are as presented in Table 1 in terms of the number of attributes, records, and classes. Six data sets were selected from the UCI repository; they were selected from different fields where data is usually limited. These datasets have about 102944–11000000 instances, and about 18–116 features. Each of the datasets has its features and parameters that are specific to it. The datasets are either numerical or text-numerical.

Table 1. Datasets description

Data No.	Name	No of record	No of attributes	No of classes
DS1	Dota2	102944	116	2
DS2	Covtype	581012	54	7
DS3	Covtype-2	581012	54	2
DS4	SUSY	5,000,000	18	2
DS5	Botnet Attacks	7,062,606	115	10
DS6	Higgs	11,000,000	28	2

The performance of the proposed models was evaluated in terms of the accuracy, sensitivity, specificity, and ROC curve. These performance measures were calculated using the expressions below:

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \tag{1}$$

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \tag{2}$$

$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP}) \tag{3}$$

The ROC curve is an important metric for performance evaluation of new models; it is produced by plotting the TPR on the y-axis & FPR on the x-axis; note that TPR means the true positive rate while FPR is the false positive rate. Both metrics are calculated as shown below:

$$\text{TPR} = \text{TP}/(\text{TP}+\text{FN}) \tag{4}$$

$$\text{FPR} = \text{FP}/(\text{TN}+\text{FP}) \tag{5}$$

True positives are the instances that were correctly predicted by the classifier as positive, while true negatives are instances correctly predicted by the classifier as negative. False positives are instances that are wrongly classified as positive by the model while false negatives are instances wrongly classified by the classifier as negative. For any prediction model to be considered good, it must achieve a point in the upper left quadrant of the ROC space that represents absolute sensitivity (meaning devoid of FPs).

6. CONCLUSION

This work aims to demonstrate the feasibility of using LDA as an attribute reduction method for big data classification. The performance of the classifier was evaluated before and after the implementation of LDA as a technique for dimensionality reduction. The classifier was tested on six different datasets pulled from the UCI ML repository and from the results, the performance of the classifiers was better after LDA has been applied to the dataset. So, LDA can be applied to linear data as a dimensionality reduction method. It can be used to improve the performance of some of the available classification models. However, the challenge of this method is that the difference between some results in some cases was much before and after applying LDA while the difference is small in some other cases. So, the application of LDA is not a sure way of drastically improving the performance of classifiers as the performance is based on the dataset employed.

Table 2. Classification Evaluation without LDA

Dataset	Classifier	Accuracy	Sensitivity	Specificity
DS1	Stochastic Gradient Descent Algorithm	0.9045	0.8089	0.9038
	Kernel Approximation Algorithm	0.8987	0.8930	0.8184
	J48	0.8732	0.8531	0.7884
	Multilayer perceptron	0.9205	0.7933	0.8642
DS2	Stochastic Gradient Descent Algorithm	0.7963	0.9445	0.9489
	Kernel Approximation Algorithm	0.7734	0.8370	0.9414
	J48	0.8043	0.9832	0.9399
	Multilayer perceptron	0.7687	0.8887	0.9332
DS3	Stochastic Gradient Descent Algorithm	0.8345	0.9191	0.8717
	Kernel Approximation Algorithm	0.7372	0.9726	0.9058
	J48	0.7453	0.8804	0.8409
	Multilayer perceptron	0.7789	0.8838	0.8875
DS4	Stochastic Gradient Descent Algorithm	0.8934	0.8787	0.8594
	Kernel Approximation Algorithm	0.9057	0.9269	0.8530
	J48	0.9003	0.8119	0.9382
	Multilayer perceptron	0.9446	0.7920	0.8788
DS5	Stochastic Gradient Descent Algorithm	0.8778	0.9357	0.9370
	Kernel Approximation Algorithm	0.9753	0.8980	0.8499
	J48	0.8380	0.9029	0.8643
	Multilayer perceptron	0.9015	0.8935	0.9566
DS6	Stochastic Gradient Descent Algorithm	0.8439	0.8367	0.8971
	Kernel Approximation Algorithm	0.9831	0.8297	0.8852
	J48	0.9928	0.8099	0.7917
	Multilayer perceptron	0.8073	0.8542	0.9621

Table 3. Classification Evaluation with LDA

Dataset	Classifier	Accuracy	Sensitivity	Specificity
DS1	Stochastic Gradient Descent Algorithm	0.9718	0.9434	0.8788
	Kernel Approximation Algorithm	0.9678	0.9933	0.9072
	J48	0.9061	0.9168	0.9241
	Multilayer perceptron	0.9097	0.8988	0.8983
DS2	Stochastic Gradient Descent Algorithm	0.8884	0.9102	0.8475
	Kernel Approximation Algorithm	0.9047	0.9334	0.9051
	J48	0.8973	0.9746	0.8905
	Multilayer perceptron	0.9487	0.9454	0.9183
DS3	Stochastic Gradient Descent Algorithm	0.8995	0.8926	0.8595
	Kernel Approximation Algorithm	0.9663	0.9179	0.9048
	J48	0.9115	0.8906	0.8997
	Multilayer perceptron	0.9988	0.9892	0.8900
DS4	Stochastic Gradient Descent Algorithm	0.9447	0.9667	0.8907
	Kernel Approximation Algorithm	0.8980	0.9233	0.9150
	J48	0.9736	0.9394	0.8905
	Multilayer perceptron	0.9344	0.9280	0.8612
DS5	Stochastic Gradient Descent Algorithm	0.9251	0.9831	0.8687
	Kernel Approximation Algorithm	0.9384	0.9326	0.8569
	J48	0.9165	0.9743	0.8953
	Multilayer perceptron	0.9661	0.9849	0.8946
DS6	Stochastic Gradient Descent Algorithm	0.9316	0.9767	0.8897
	Kernel Approximation Algorithm	0.9935	0.9021	0.8549
	J48	0.9101	0.9289	0.9124
	Multilayer perceptron	0.9904	0.9401	0.9329

ABBREVIATIONS:

LDA: Linear Discriminant Analysis

CONFLICT OF INTERESTS

None to declare.

FUNDING

No funding was received for this work.

ACKNOWLEDGMENTS

The moral support from the Iraqia University and Al Salam University College is appreciated.

REFERENCES

[1] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering*, vol. 11, no. 6, pp. 138–150, 2019.
[2] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," *In 2018 1st Annual International Conference on Information and Sciences (AiCIS)*, pp. 140–145, 2018. IEEE.
[3] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.
[4] J. Wen *et al.*, "Robust sparse linear discriminant analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 390–403, 2018.
[5] S. Ji and J. Ye, "Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection," *IEEE Transactions on Neural Networks*, vol. 19, no. 10, pp. 1768–1782, 2008.
[6] D. Chu, L. Z. Liao, M. K, . P. Ng, and X. Wang, "Incremental linear discriminant analysis: A fast algorithm and comparisons," *IEEE transactions on neural networks and learning systems*, vol. 26, pp. 2716–2735, 2015.
[7] J. Ghosh and S. B. Shuvo, "Improving Classification Model's Performance Using Linear Discriminant Analysis on Linear Data," *In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, 2019. IEEE.
[8] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982–994, 2004.
[9] M. Zhang, R. Chen, X. Zhang, Z. Feng, G. Rao, and X. Wang, "Intelligent rdd management for high performance in-memory computing in spark," *in Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 873–874, 2017.
[10] R. A. Hasan, R. A. I. Alhayali, N. D. Zaki, and A. H. Ali, "An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark," *Telkomnika*, vol. 17, no. 6, pp. 3086–3099, 2019.
[11] E. Alomari, I. Katib, and R. Mehmood, "Iktishaf: A big data road-traffic event detection tool using Twitter and spark machine learning," *Mobile Networks and Applications*, pp. 1–16, 2020.
[12] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 4, pp. 1411–1419, 2020.
[13] A. H. Ali and M. Z. Abdullah, "A novel approach for big data classification based on hybrid parallel dimensionality reduction using spark cluster," *Computer Science*, vol. 20, no. 4, 2019.
[14] A. H. Ali and M. Z. Abdullah, "A Parallel Grid Optimization of SVM Hyperparameter for Big Data Classification using Spark Radoop," *Karbala International Journal of Modern Science*, vol. 6, no. 1, p. 3, 2020.
[15] F. Nie, Z. Wang, R. Wang, Z. Wang, and X. Li, "Adaptive local linear discriminant analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 14, no. 1, pp. 1–19, 2020.
[16] H. Li, L. Zhang, B. Huang, and X. Zhou, "Cost-sensitive dual-bidirectional linear discriminant analysis," *Information Sciences*, vol. 510, pp. 283–303, 2020.
[17] F. Zhou and G. Cong, "On the convergence properties of a $ K $-step averaging stochastic gradient descent algorithm for nonconvex optimization," 2017. arXiv preprint arXiv:1708.01012.
[18] O. A. Arqub and B. Maayah, "Fitted fractional reproducing kernel algorithm for the numerical solutions of ABC – Fractional Volterra integro-differential equations," *Chaos, Solitons & Fractals*, vol. 126, pp. 394–402, 2019.
[19] H. Hong, "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)," *Catena*, vol. 163, pp. 399–413, 2018.
[20] Z. Car, S. B. Šegota, N. Anđelić, I. Lorencin, and V. Mrzljak, "Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron," *Computational and Mathematical Methods in Medicine*, vol. 2020, pp. 1–10, 2020.