INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Management, Science and Technology (ICMST 2021)

Comparison of ECB-LDA and DSO-RBNN in Diabetes Prediction using Big Data Analytics

G.GeoJenefer[1], *Dr.A.J Deepa*[2]
[1]*Research Scholar, Anna University, Chennai, TamilNadu, India*
[2]*Professor, Ponjesly College of Engineering, Nagercoil, TamilNadu, India*
geo.jenefer@gmail.com[1]

Abstract

Diabetes is one of the chronic diseases rovering all over the world. It affects people in all ages. Even child by birth also getting affected by this disease. Already various machine learning algorithms were used to predict diabetes. This work compares two algorithms Enhanced Catboost with Linear Discriminant Analysis (ECB-LDA) and Dolphin Swarm Optimization with Radial Basis Neural Network (DSO-RBNN) which were used for diabetes prediction. Also hospitals and other clinical centers are facing problem in handling large amount of data. To solve such problem and also do early prediction of diabetes, big data analytics is used. This work proves that the accuracy of DSO-RBNN is better than the ECB-LDA.

Keywords: Big Data Analytics, Chronic Disease, Linear Discriminant Analysis (LDA), Dolphin Swarm Optimization (DSO), Radial Basis Neural Network (RBNN).

1. Introduction

Nowadays, society is handling an enormous amount of data which is further said to be Big Data. Various industries are facing a lot of issues while handling such kinds of data. Big Data Analytics is a concept which analyses big data and also helps in predicting future happenings. This work focuses on predicting the disease in the health industry that generates bulk data. The sources are clinical reports, diagnostic reports, laboratory reports, doctor's prescription, medical images, pharmacy information, Electronic Health Reports, Health insurance reports, etc. Information gathered from these sources is said to be Big Data. It is inevitable to analyse and deal with this big data. A lot of chronic diseases are still in the health industry, which extends for the long term. These diseases could be cured if predicted earlier. Big Data Analytics plays a vital role in analysing the data and predicting the disease earlier.[1-5]. This work results better in predicting chronic diseases. For experimentation, the PIMA diabetes dataset is used as the input data and has obtained good results.

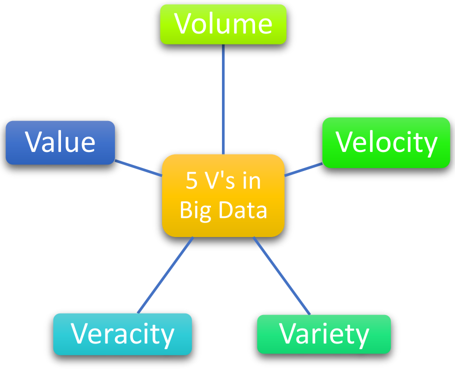

Fig 1: 5 V's in Big Data

Big data provides 5 dimensionalities (Volume, Velocity, Variety, Veracity, and Value) which are necessary for managing flooded data.

Volume

It indicates the large amount of data collected through different sources. Approximately they generate 2.5 quintillion bytes of data every day. Big Data plays an essential role in handling such voluminous data.

Velocity

It denotes the speed of the generated data. While considering social media, it uploads millions of data on Facebook, Twitter, youtube, and google daily. Big Data helps organizations to handle those data faster.

Variety

It denotes the structured and unstructured data obtained from different sources. Some of the structured data are text, pictures, video, etc., and unstructured videos are emails, voice mails, audio recordings, etc.

Veracity

It refers to the quality of the data. Before processing the data, it should analyse whether the data is clean and accurate to improve the quality.

Value

It analyses the value of the data and converts the bulk amount of data into business.

1.1 Machine Learning

Machine learning is a recent innovation technology which helps mankind in improving many industrial and professional process in daily life.

Various types of machine learning algorithms are:

•**Supervised:** In supervised learning, the algorithms are trained using labelled examples.

•**Unsupervised:** In unsupervised learning, the algorithm uses data that has no historical labels.

•**Semi-supervised:** It uses both labelled and unlabeled data for training.

•**Reinforcement Learning:** Used in robotics, gaming, etc.

1.2 BDA In Healthcare

Health care faces a lot of challenges while handling an enormous amount of data. For handling such complex datasets and making more efficient decisions, BDA is used. BDA retrieves patient’s information from Electronic Health Record data, Electronic Medical Record data, imaging data, or Sensor data. Then convert it into information necessary to the doctors or researchers, or analysts to make proper decisions. It helps the health care industry by providing personalized medicine through prescriptive analytics, predictive analytics, etc. This survey focuses on predicting chronic diseases.

A single patient may have files such as electronic health records, doctor prescription, lab results, insurance, medical equipment, etc. It is impossible to analyze such kind of dataset. BDA helps to analyze such a dataset.

1.3 Chronic Disease

Chronic disease is a disease that persists for a long time. Such disease can be caused by the usage of tobacco, lack of physical activity, poor eating habit, etc. Some examples are heart disease, diabetes, kidney failure, cancer, stroke, arthritis, asthma, ulcer, obesity, etc. This paper used diabetic patients’ dataset and did early prediction of the disease. Diabetes is one of the deadly diseases which must be predicted earlier to decrease the severity of the disease. Such prediction may also help the medical practitioners to make better decision before giving treatment.

1.4 Spark

Spark is an open source framework used in this work. It has the facility to add more features and efficiencies with the existing software. It is used in large companies for handling huge data. It integrates with different file system such as HDFS, MONGODB and amazon’s s3 system since it doesn’t have its own filesystem.

Spark is a leading platform for large scale SQL, batch processing, stream processing and machine learning.

Features:

•It requires large memory for processing with huge amount of data.

•It supports most of the programming languages.

•Scalability and fault tolerance.

2. Research Studies

Sneha et al., experimented Decision Tree, Naïve Bayes and Random Forest algorithms. From these it was found that Decision and Random Forest obtained highest specificity of 98.2% and 98% respectively. While comparing the accuracy Naïve Bayes obtained the first with 82.3%. Lai et al., compared various Machine Learning algorithms and found that Gradient Boosting Method and Logistic Regression obtained sensitivity of 71.6% and 73.4% respectively. These algorithms perform better than Random Forest and Decision Tree(DT) Models. Quan Zou et al., used Minimum

Redundancy Maximum Relevance(MRMR) algorithm for feature selection which works better than Principal Component Analysis (PCA). Random Forest along with MRMR feature selection algorithm works better than DT and Neural Networks. The performance of this work with PIMA results with 0.7721 whereas with Luzhou dataset its 0.8084. Alam et al., implemented Artificial Neural Network (ANN), Random Forest (RF) and K-means clustering algorithms. From these, ANN provides best accuracy of 75.7%. This work was implemented in PIMA dataset which contain 768 records with 9 attributes. The research gap found in this work was it can be used only for structured dataset. Larabi-Marie-Sainte et al., combined (Convolution Neural Network) CNN with LTSM (Long Term Short Memory) and achieved 95.1% accuracy. They surveyed all ML and DL(Deep Learning) techniques implemented in the last six years and found that the accuracy for ML ranges from 68% to 74%. The highest accuracy of DL is 95%. Xue et al., used 520 patients diabetic dataset aged between 16 to 90. They implemented various algorithms like SVM(Support Vector Machine), NB and light GBM and proved that SVM performs better when comparing with other algorithms. Mujumdar et al., compared various machine learning algorithms in predicting diabetes dataset and found that Logistic Regression secured accuracy of 96%. After applying pipeline in those algorithms, Adaboost classifier performed better than LR and obtained 98.8% accuracy. Sisodia et al., compared performance metrices like precision, accuracy, F-measure and recall of various algorithms like DT, SVM and NB. They found that NB performs and well and obtained 76.3% accuracy. From the above research studies, it was analysed that various ML and DL algorithms were already implemented to predict diabetes. Further studies should concentrate on various combination of algorithms to improve performance measures and also these should be implemented in unstructured data.[4-9].

3. Methodology

This work uses ECB-LDA and DSO-RBNN algorithms in PIMA diabetes dataset.

3.1 ECB-LDA

Pseudocode

1. ECB-LDA takes the input data from PIMA diabetic dataset.
2. Pre-process the input data, so that the data is split as training and test data.
3. Apply min-max feature scaling to normalize the input variables. The variable ranges between 0 and 1.

Min-Max feature scaling is calculated as,

$$x1 = x - \frac{\min(x)}{\max(x)} - \min(x) \qquad (1)$$

Where x denotes the original value and x1 denotes the normalized value.

Min-Max Normalization is calculated as,

$$x1 = a + (x - \min(x))(b - a)/\max(x) - \min(x)(2)$$

a, b – min and max values.

Mean Normalization is calculated as

$$x1 = x - \frac{average\ (x)}{\max(x)} - \min(x) \qquad (3)$$

4. After feature scaling, the dimensionality of the dataset is reduced using LDA.

3.2 DSO-RBNN

Steps included:

1. Preprocess the input data for choosing the accurate attributes.
2. Fast Correlation Based Feature Selection technique is used to reduce the dimensionality and to obtain accurate value in prediction.
3. Then, Clustering is done using Hadoop K-means clustering technique.
4. The dataset is divided into number of clusters using DSO, an optimization algorithm.
5. Finally, Classification is done using RBNN to classify diabetic and non-diabetic patients.

4. Performance Metrics

Precision:

Precision defines the number of accurate predictions done. It is the proportion of true positive with true positive and true negative. Precision of ECB-LDA is 0.99 and DSO-RBNN is 0.89. It was proven that precision of ECB-LDA for PIMA diabetic dataset is higher.

Precision is measured using,

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (4)$$

Where, TP indicates True Positive and FP indicates False Positive.

Recall

It defines the proportion of predicted True Positives from all True positives.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (5)$$

Where, TP indicates True Positive and FN indicates False Negative.

Accuracy

It denotes the number of correct predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

Where, TP indicates True Positive, TN indicates True Negative, FP indicates False Positive and FN indicates False Negative.

Table 1: Performance metrices of ECB-LDA and DSO-RBNN

Sl.No.	Algorithm	Precision	Recall	Accuracy
1	ECB-LDA	0.99	0.9	0.96
2	DSO-RBNN	0.89	0.91	0.98

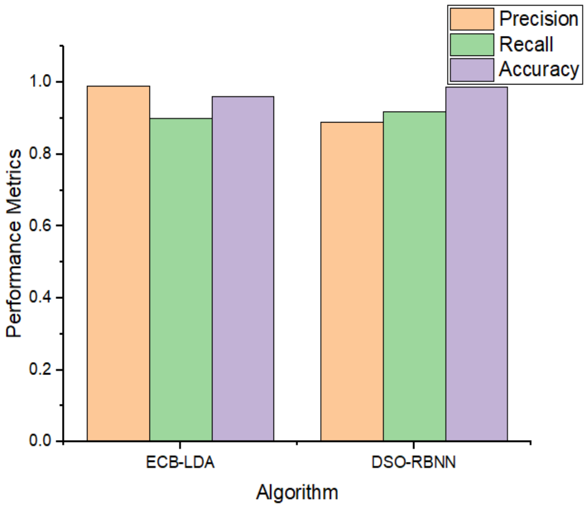

Fig.1: Comparison of performance metrices with ECB-LDA and DSO-RBNN algorithms

Based on the comparison of precision, recall and accuracy of ECB-LDA with DSO-RBNN it was proven that DSO-RBNN response better for recall and accuracy with good percentage.

Conclusion

Diabetes is a deadly chronic disease which must be predicted earlier and treated as well. Various ML and DL algorithms were already predicted. But still those algorithms face performance issues. This work compared two algorithms and the performance metrices were discussed. It was proven that the accuracy of DSO-RBNN performs better than ECB-LDA. In future, this algorithm must be used with different disease datasets like heart disease, cancer, COVID19 etc.

References

Journals

[1]. Sneha, N., and TarunGangil. "Analysis of diabetes mellitus for early prediction using optimal features selection." Journal of Big data 6, no. 1 (2019): 1-19.

[2]. Lai, Hang, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. "Predictive models for diabetes mellitus using machine learning techniques." BMC endocrine disorders 19, no. 1 (2019): 1-9.

[3]. Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[4]. Alam, Talha Mahboob, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain et al. "A model for early prediction of diabetes." Informatics in Medicine Unlocked 16 (2019): 100204.

[5]. Larabi-Marie-Sainte, Souad, LinahAburahmah, Rana Almohaini, and Tanzila Saba. "Current techniques for diabetes prediction: review and case study." Applied Sciences 9, no. 21 (2019): 4604.

[6]. Xue, Jingyu, Fanchao Min, and Fengying Ma. "Research on Diabetes Prediction Method Based on Machine Learning." In Journal of Physics: Conference Series, vol. 1684, no. 1, p. 012062. IOP Publishing, 2020.

[7]. Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[8]. Berbudi, Afiat, NofriRahmadika, Adi Imam Tjahjadi, and RovinaRuslami. "Type 2 diabetes and its impact on the immune system." Current diabetes reviews 16, no. 5 (2020): 442.

[9]. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.