ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter160892020.pdf
https://doi.org/10.30534/ijeter/2020/160892020

Hybrid Dimensionality Reduction for Outlier Detection in High Dimensional Data

Roy Thomas[1], **J.E.Judith**[2]
[1] Noorul Islam Centre for Higher Education, Kumaracoil, India, roygptc@gmail.com
[2]Noorul Islam Centre for Higher Education, Kumaracoil,India. judithjegan@gmail.com

ABSTRACT

The infrequent data objects in a dataset containing important characteristics are called outliers. Dimensionality reduction techniques are used for reducing the memory requirement and processing time while handling high dimensional datasets. Principal Component Analysis (PCA) is one of the most commonly used techniques for reducing the dimensions in high dimensional datasets. Autoencoders (AE) are used in deep learning for dimensionality reduction. In this paper, we evaluate the performance of these dimensionality reduction techniques for outlier detection task and also propose a hybrid dimensionality reduction technique that can be used in the preprocessing step of many machine learning models. Experimental results using publicly available datasets show that the proposed method outperforms the PCA method and autoencoder based method in outlier detection tasks using isolation forest.

Key words : Autoencoder, Dimensionality reduction, Isolation Forest, Principal component.

1. INTRODUCTION

Data mining and machine learning tasks require the processing of datasets containing large volume of data objects having different characteristics. A number of features are required to completely describe each data object in the dataset. These features can be represented using numerical or categorical variables which form different dimensions of the dataset. The processing of high dimensional datasets requires a large volume of computer memory and processor time. Data preprocessing is an initial step in many data mining and machine learning tasks, which involves data reduction in addition to data cleaning, integration, discretization, transformation etc.

The objective of data reduction task in machine learning is to construct a reduced representation of the original dataset. The reduced representation will contain less volume of data, but will produce almost the same result as produced by the original dataset. Different strategies are used in the literature for data reduction such as data cube aggregation, data compression, dimensionality reduction and concept hierarchy generation. Dimensionality reduction may be defined as the process of reducing the number of dimensions of the dataset. The original dataset may contain a number of columns where each column represents a particular feature of the dataset. The main objective of dimensionality reduction is to bring the number of dimensions to a manageable size without compromising the result from the data mining or machine learning model.

There are some difficulties in processing high dimensional datasets which do not exist in lower dimensional datasets. The term ‘curse of dimensionality’ is used to denote the problems that arise while processing high dimensional datasets. An increase in number of features will cause an increase in the number of combinations of the features to be represented by the machine learning model. This makes the model more complex and increases the chance of overfitting, which in turn results in poor performance. If the number of features is less, the model will be simple and the problem of overfitting is avoided. In addition to avoiding overfitting, dimensionality reduction helps to improve the model accuracy, to effectively utilize the memory and to process the algorithms faster. Certain data mining and machine learning algorithms which are unfit for high dimensional datasets can be used for processing such large datasets after reducing the dimensions to an affordable dimension by dimensionality reduction techniques.

In this paper, outlier detection technique is used to evaluate and compare the performance of dimensionality reduction techniques. “Outliers are data objects in a dataset whose characteristics are not in accordance with the characteristics of the majority of the data objects in the dataset”[1].Outlier detection is an important problem in statistics as well as in data mining and machine learning, since it has a wide range of applications such as fraud detection in financial transactions, fault diagnosis in industrial products, disease identification in medical field, intrusion detection in network traffic etc. Outlier detection algorithms can be categorized into different categories. Distance based outlier detection methods, clustering based methods, classification based methods and statistical methods are some among them. In each category, different algorithms are developed by the researchers depending on the nature of the datasets. We, in this paper, use the isolation forest model for evaluating the performance of the proposed hybrid dimensionality reduction technique.

2. RELATED WORK

Outlier detection techniques and dimensionality reduction techniques have been an important area of research for many years. Different techniques have been developed for dimensionality reduction using non-linear transformations as well as linear transformations. Jolliffe [2] explained the popular dimensionality reduction method 'Principal Component Analysis' with proof. Most of the algorithms used for finding anomaly detection can also be used for detecting outliers with minor modifications. Hodge et al. [3] accumulated the outlier detection techniques in statistics, neural networks and machine learning. They classified the outlier detection techniques into three types depending on the availability of class labels.

Chandola, Banerjee and Kumar[4] accumulated the outlier detection methods into different groups such as statistical based, classification based, clustering based, nearest neighbor based, spectral methods and information theoretic methods. In addition to the detection of global outliers, Han, Kamber and Pei [5] presented methods for detecting collective outliers and contextual outliers. Chalapathy and Chawla [6] accumulated different anomaly detection methods and presented them in the perspective of deep learning. They described the advantages and disadvantages of deep learning based anomaly detection techniques and also explained the application areas of various methods. Dai, Yan, Wang and Zhang [7] proposed a one class model for anomaly detection using autoencoders and support vector machine model. Thomas and Judith[8] proposed voting based ensemble of outlier detectors to improve the performance of individual outlier detection algorithms. In [9], Putri et al. made a comparative study of three outlier detection algorithms for time series data. Niranjan et al.[10] Proposed a ensemble model for efficient classification of phishing using different machine learning algorithms. In [11], Govindarajan proposed an ensemble method that employs Naïve Bayes ,Support Vector Machine) and Genetic Algorithm as base classifiers for text categorization. Methods for detecting outliers in datasets containing categorical features [12] have also been proposed by many researchers.

3. PROPOSED MODEL

This paper presents a hybrid model for dimensionality reduction that can be applied for detecting outliers in high dimensional datasets. The architecture of the model is shown in Figure.1. The following subsections explain the different phases in the model.

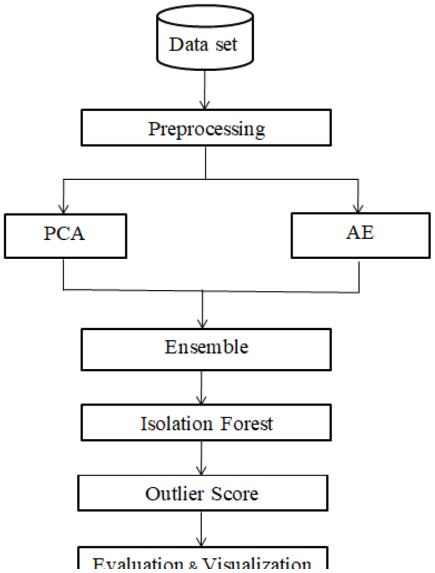

Figure 1: Hybrid model

3.1 Preprocessing

Data preprocessing is the initial step in the proposed model. The raw data may contain some error data and missing data. For our experiments, we use publicly available datasets from the UCI machine learning repository. The experiments aim to evaluate the effectiveness of the model for detecting outliers in high dimensional datasets. The percentage of outliers in a dataset is very less compared to the percentage of normal data objects in the dataset. The strategy for handling the missing data in the proposed model is to avoid the records containing the missing values, since replacing them with some calculated values may severely affect the percentage of outliers in the dataset.

3.2 Classification of Dimensionality Reduction Techniques

Feature selection methods and feature extraction methods are the two main categories of dimensionality reduction techniques. In feature selection methods, relevant features from the set of features are identified and selected for processing. Feature extraction methods deal with the generation of new features by applying some transformation to the existing features. The widely used dimensionality reduction methods are based on the application of linear transformations to the high dimensional data. Some of the algorithms that use linear transformations for dimensionality reduction are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Factor analysis. If the data is not in a linear subspace, non-linear transformation or manifold learning methods are used to reduce the dimensions, which are based on the hypothesis that in a high dimensional feature space, a small number of low dimensional manifolds will contain the most relevant data. Multi-Dimensional Scaling (MDS) and Isometric feature mapping are examples of dimensionality reduction techniques that use non-linear transformation. In this paper, a model that ensemble a linear transformation method and a non-linear transformation method is proposed. Principal component analysis and autoencoders are used for this purpose.

3.3 Principal Component Analysis

PCA is one of the most widely used dimensionality reduction techniques for high dimensional datasets containing continuous data.PCA projects data along the direction of increasing variance and principal components are the features with the maximum variance. As an example, Figure.2 shows the scatter plot of two features x1 and x2 in x and y directions respectively. Here the dashed line contains instances that vary the most. PCA calculates the eigenvalues and eigenvectors of the covariance matrix for the variables in the standardized dataset [2]. Then k (the reduced dimension) eigenvalues are taken from the sorted list of eigenvalues and the corresponding eigenvectors to form a matrix of eigenvectors. Using this matrix of eigenvectors, the original matrix of the dataset is transformed to k-dimensional space.

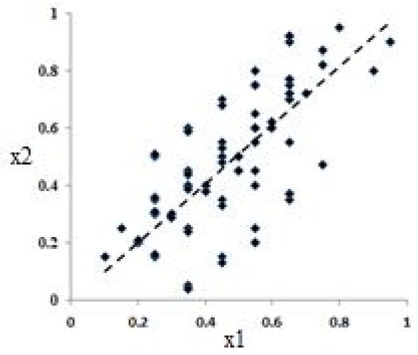

Figure 2: Principal component

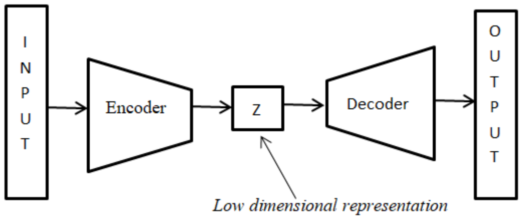

Figure 3:Autoencoder

3.4 Autoencoders

Autoencoders are also used for reducing the dimensions of high dimensional data in deep learning and they give better results than conventional dimensionality reduction methods. Autoencoder is a type of artificial neural network that compresses the input into a latent space representation and reconstructs it from this latent space representation. Basically an autoencoder is a type of unsupervised learning structure which consists of an input layer, one or more hidden layers and an output layer. An autoencoder can be considered as a combination of an encoder followed by a decoder [4]. The encoder part is responsible for transforming the high dimensional input data into a low-dimensional representation. The encoder output is given as input to the decoder and the decoder reconstructs the original data from this low dimensional representation. Figure.3 illustrates the operation of an autoencoder.

3.5 Hybrid Dimensionality Reduction

In this method, we use an ensemble technique for dimensionality reduction. The high dimensional input dataset is given to PCA and autoencoder models separately. The results obtained from these dimensionality reduction techniques are normalized and combined using the ensemble learning technique. The ensemble learning approach used in this paper is a weighted averaging method as given in (1), where P_1 and $P_2$are the outputs of the dimensionality reduction techniques used in the model, $w_1$and w_2 are the corresponding weights assigned to them and P is the result obtained from this ensemble method.

$$P = (w_1P_1+w_2P_2)/(w_1+w_2) \qquad (1)$$

There are different methods for normalizing the data before using in any data mining or machine learning model. In this model min-max scaling is used for bringing the outputs of the dimensionality reduction methods into a standard scale, as min-max scaling is found to give better results in outlier detection methods. Figure.1 illustrates the method incorporated in the hybrid dimensionality reduction method.

3.6 Outlier Detection

Outlier detection method used in this approach is a model for an unbalanced classification problem which contains only two classes of data. This is an unbalanced binary classification problem in which the major class contains the normal objects and the minor class contains the outliers. Different types of algorithms such as neighbourhood based, Bayesian networks, decision tree based, information theoretic, clustering based and statistical techniques are used in the literature for finding outliers[4][5]. The outlier detection model used in this paper is a decision tree based algorithm called isolation forest.

3.7 Isolation Forest

Isolation forest is used in this paper for detecting outliers due to its ability to explicitly identify the outliers instead of ascertaining the normal data objects [13]. Like other decision tree based algorithms, it partitions the dataset by selecting a random feature and then randomly selecting a split value between the minimum and maximum values of this feature. Since outliers are rare in the dataset compared to the majority normal data objects, they can be found closer to the root of the tree. The average path length can be used as a score to determine the outlier degree in decision tree based approaches. The equation for calculating the outlier score in isolation forest is given in (2) where, x is the data object with n external nodes, h(x) is the path length of x, and c(n) is the average path length of unsuccessful search in binary search tree.

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{2}$$

The score obtained is used to check whether a data object is an outlier or not. A value very close to 1 indicates that the data object is an outlier. A threshold value is used to compare the outlier score, so that data objects having outlier score greater than the threshold are taken as outliers and objects having outlier score less than or equal to the threshold are taken as inliers.

4. EXPERIMENTS AND RESULTS

This section describes the experiments using the proposed model and the results obtained from the experiments. There are two subsections in this section. The first subsection provides the experimental settings and datasets. Results and performance evaluation are given as another subsection.

4.1 Experimental Settings and Datasets

Python language was used to conduct experiments in an Intel core i5-based computer. Public datasets from the UCI machine learning repository were used for the experiments. One of the datasets used for the experiment was the breast cancer dataset[14]. The original dataset contained 569 instances and 32 attributes. Out of these 32 attributes, the id attribute was removed as it would not give any contribution to the classification algorithms. Another attribute was the class attributes having the values malignant and benign. In order to make the percentage of outliers less, we took 300 instances of normal data objects and 26 instances of outlier instances. Thus the input dataset used for the experiment contained 326 instances and 30 attributes.

Another dataset used for the experiment was cardiotocography dataset [15]. The characteristics of these datasets are described in Table 1. The performance of the various dimensionality reduction techniques PCA, autoencoder and hybrid model was evaluated by the efficiency of their outputs in finding outliers. The outlier detection algorithm used in the experiment was the isolation forest. The results obtained from these different models were compared with the actual dataset and evaluated the performance of the models.

Table 1: **.**Dataset caharacteristics

Dataset	Breast Cancer	Cardiotocography
#instances	569	2126
#attributes	32	23
Attribute type	Real	Real
#classes	2	3
#normal	300	1655
#outlier	26	166

4.2 Results and Evaluation

Figure 4 shows the scatter plot of the results obtained from the various dimensionality reduction techniques when applied to the breast cancer dataset for detecting the outliers. The results shown in the plot are the predicted result of the outliers after applying the dimensionality reduction technique. In order to show the plot in two-dimensional space, we have reduced the dimension of the high dimensional data from thirty to two, namely x1 and x2, which are taken in x and y axes respectively. This scatter plot gives a visual appearance of the performance of dimensionality reduction techniques in which the proposed hybrid method gives better performance.

F1-score is a good measure for evaluating the performance of the outlier detection algorithms. In this paper we used F1-score and accuracy for comparing the performance of the individual and hybrid models for dimensionality reduction. The formulae for computing these metrics are derived from true positive (TP), false positive (FP) , true negative (TN) and false negative (FN). Precision (3) and recall (4) are used to compute the F1-score (5). Accuracy (6) is the fraction of correct predictions from all the predictions.

$$Precision = \frac{TP}{(TP+FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{4}$$

$$F1\_\text{Score} = 2 * \frac{Precision * Recall}{(Precision + Recall)} \tag{5}$$

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

The outlier class is taken as the positive class as the experiment was conducted to detect the outliers and the normal class is the negative class. The values of these evaluation measures obtained for various models using the breast cancer dataset are tabulated in Table 2. Table 3 shows these values for the cardiotocography dataset.

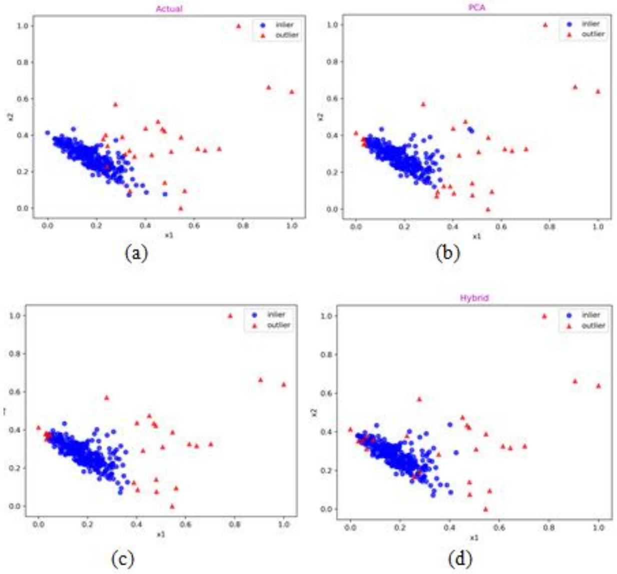

Figure 4: Scatter plot of outliers(a) Actual (b)PCA(c) Autoencoder (d)Hybrid

Table 2. Evaluation result for breast cancer dataset

Dimension	Measure	PCA	Autoencoder	Hybrid (Proposed)
24	F1-score	0.423	0.346	**0.538**
	Accuracy	0.908	0.896	**0.926**
14	F1-score	0.423	0.577	**0.615**
	Accuracy	0.908	0.933	**0.939**
4	F1-score	0.500	0.538	**0.692**
	Accuracy	0.920	0.926	**0.951**

Table 3: Evaluation result for cardiotocography dataset

Dimension	Measure	PCA	Autoencoder	Hybrid (Proposed)
16	F1-score	0.464	0.422	**0.536**
	Accuracy	0.902	0.895	**0.915**
9	F1-score	0.536	0.482	**0.633**
	Accuracy	0.915	0.906	**0.933**
3	F1-score	0.506	0.518	**0.536**
	Accuracy	0.910	0.912	**0.915**

Figure5 and Figure6 show the F1-score and accuracy for different dimensions as a graph for the breast cancer dataset. It is concluded from the graph that the proposed hybrid model for dimensionality reduction outperforms the Principal Component Analysis and autoencoder models. Figure 7 and Figure 8 show these measures for the cardiotocography dataset corresponding to various dimensions. Here also, our proposed hybrid model outperforms the others.

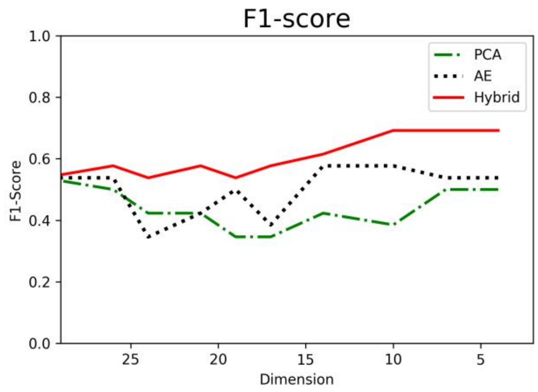

Figure5.F1-score (breast cancer dataset)

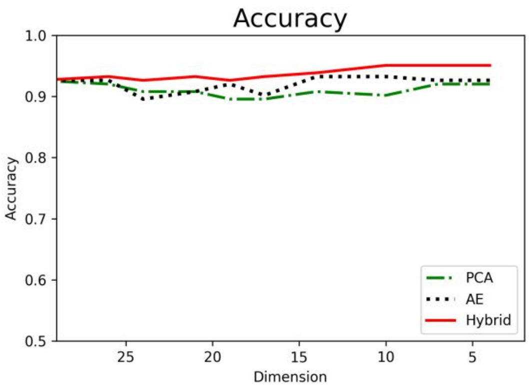

Figure 6: Accuracy (breast cancer dataset)

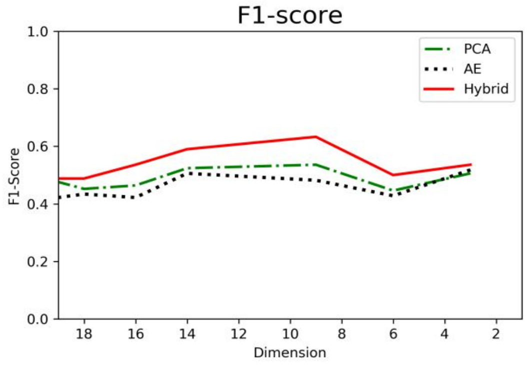

Figure 7 : F1-score (cardiotocography dataset)

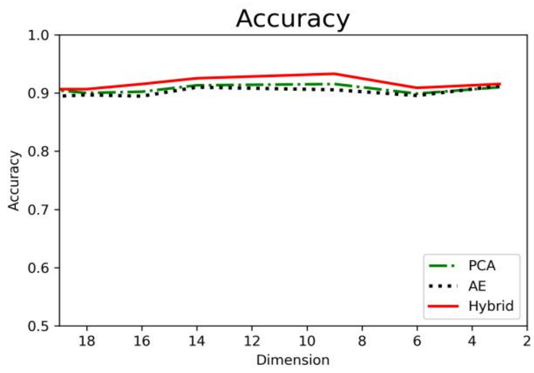

Figure 8: Accuracy (cardiotocography dataset)

5. CONCLUSION

In this paper we analyzed and evaluated the performance of the dimensionality reduction techniques PCA and autoencoder individually and compared the performance of these techniques with the performance of the proposed hybrid model. The machine learning task used for evaluating the proposed model was the outlier detection method using isolation forest. We studied the performance using the publicly available breast cancer dataset and we were able to achieve higher F1-score and accuracy for our proposed method which shows that the proposed method outperforms the existing methods. The aim of the method was to study the performance of the dimensionality reduction methods on unbalanced classification problems. The proposed method can be used for any binary class classification problems.

REFERENCES

1. D. Hawkins. **Identification of Outliers**. *Chapman and Hall,* London and New York, 1980.
2. I. T. Jolliffe. **Principal component analysis**, *Springe*r,2002.
3. V. J. Hodge, and J. Austin. **A survey of outlier detection methodologies**, *Artificial Intelligence Review,* vol. 22 (2), pp. 85-126, 2004.
4. V. Chandola,, A. Banerjee, and V. Kumar. **Anomaly detection: A survey**,*ACM Comput. Surveys*, vol. 41, no. 3, pp. 1–58, 2009.
5. J. Han, M. Kamber, and J. Pei,.**Data Mining: Concepts and Techniques**, Massachusetts (US): Morgan Kaufmann, 2012.
6. R. Chalapathy, and S. Chawla.**Deep learning for anomaly detection: a survey**, arXiv:1901.03407, 2019 [online] Available: https://arxiv.org/abs/1901.03407
7. S. Dai, T. Yan, X. Wang, and L. Zhang. **A Deep One-class Model for Network Anomaly Detection**,*IOP Conf. Series: Materials Science and Engineering,* **563 (2019)** 042007.
8. R. Thomas, and J. E. Judith. **Voting-Based Ensemble of Unsupervised Outlier Detectors**, *Advances in Communication Systems and Networks*, pp. 501-511. Springer, Singapore, 2020.
9. D. A. P. Putri, and E. Sudarmilah. **Comparative Study for Outlier Detection In Air Quality Data Set**, *International Journal of Emerging Trends in Engineering Research*,Vol.7(11),pp.584-592, 2019.
10. V.Niranjan,. K. Sakhamuri ,P. D.Shenoy,and K. R. Venugopal .**ERCRFS: Ensemble of Random Committee and Random Forest using StackingC for Phishing Classification,** *International Journal of Emerging Trends in Engineering Research*, Vol.8(1),pp.79-86, 2020.
11. M. Govindarajan. **Ensemble of Classifiers in Text Categorization,** *International Journal of Emerging Trends in Engineering Research*, Vol.8(1),pp.41-45, 2020.
12. R. Thomas, and J. E. Judith. **A Novel Ensemble Method for Detecting Outliers in Categorical Data,***International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9 (4), pp. 4947 – 4953,2020.
13. F. T. Liu, K. M. Ting, and Z. Zhou. **Isolation Forest**, *Eighth IEEE International Conference on Data Mining*, Pisa, 2008, pp. 413-422.
14. W. H. Wolberg, W. N. Street, and O.L. Mangasarian**. UCI Machine Learning Repository,** [online] Available: https://archive.ics.uci.edu/ml/datasets/Breast+ Cancer+Wisconsin+(Diagnostic)
15. M. SÃ¡, J. P. Bernarde, and J.A. Campos. **UCI Machine Learning Repository**, [online] Available: http://archive.ics.uci.edu/ml /datasets/Cardiotocography.