INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 14, Issuse.2 April 2023

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 14, Issuse.2 April 2023

Design and Applying Machine Learning Algorithms to Predict the Anomaly in Real Time

MOHAMMED AKRAM, DR. N. SUDHAKAR YADAV,

Abstract:

Anomaly detection has been used for decades to identify and extract anomalous components from data. Many techniques have been used to detect anomalies. One of the increasingly significant techniques is Machine Learning (ML), which plays an important role in this area. In this research paper, we conduct a Systematic Literature Review (SLR) which analyzes ML models that detect anomalies in their application. Our review analyzes the models from four perspectives; the applications of anomaly detection, ML techniques, performance metrics for ML models, and the classification of anomaly detection. In our review, we have identified 290 research articles, written from 2000-2020, that discuss ML techniques for anomaly detection. After analyzing the selected research articles, we present 43 different applications of anomaly detection found in the selected research articles. Moreover, we identify 29 distinct ML models used in the identification of anomalies. Finally, we present 22 different datasets that are applied in experiments on anomaly detection, as well as many other general datasets. In addition, we observe that unsupervised anomaly detection has been adopted by researchers more than other classification anomaly detection systems. Detection of anomalies using ML models is a promising area of research, and there are a lot of ML models that have been implemented by researchers. Therefore, we provide researchers with recommendations and guidelines based on this review.

Keywords: anomaly, supervised learning, machine learning.

Introduction:

Detecting anomalies is a major issue that has been studied for centuries. Numerous distinct methods have been developed and used to detect anomalies for different applications. Anomaly detection refers to ‘‘the problem of finding patterns in data that do not conform to expected behaviour’’. The detection of anomalies is widely used in a broad variety of applications. Examples of these include.

Ph.D (CSE), Assistant professor, Depertment of Information Technology. Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology (VNR VJIET), pragati nagar, Nizampet (S.O.), Hyderabad 500090

fraud detection, loan application processing, and monitoring of medical conditions, An example of a medical application is heart rate monitors. Other widely used applications of detecting anomalies include cyber security intrusion detection, fault detection for aviation safety study, streaming, and hyper-spectral imagery, etc. The importance of detecting anomalies in various application domains concerns the risk that unprop-The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Ravel . tested data may represent significant, critical, and actionable information. For instance, detecting an anomalous computer network traffic pattern may expose an attack from a hacked computer. Another example would be the detection of anomalies in the transaction data of a credit card, which may indicate theft. Besides, detecting an anomaly from an airplane sensor may result in the detection of a fault in some of the components of the aircraft.

Related work:

Modules

1. **DATA COLLECTION**
2. **DATA PRE-PROCESSING**
3. **FEATURE EXTRATION**
4. **EVALUATION MODE**

DATA COLLECTION

Data collection is a process in which information is gathered from many sources which is later used to develop the machine learning models. The data should be stored in a way that makes sense for problem. In this step the data set is converted into the understandable format which can be fed into machine learning models.
Data used in this paper is a set of cervical cancer data with 15 features . This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labelled data*.

DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

Formatting: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 14, Issuse.2 April 2023

Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance.

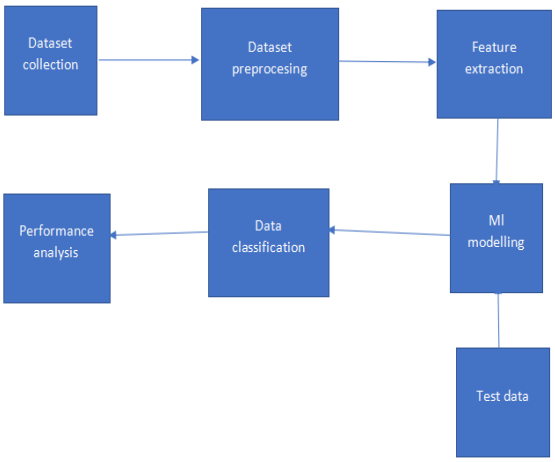
Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs.

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Proposed system:

Different supervised machine learning algorithms are implemented . such as Random forest and Logistic Regression. From these 2 algorithms we got the accuracy of logistic regression is 95% and the best result with Random forest 99.99%.

Block diagram:



Advantages

- Increasing the accuracy score
- Large amount of feature we are taking for the training and testing.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 14, Issuse.2 April 2023

System Requirements

Software:

- Anaconda Navigator
- in Python Language
- Jupyter Notebook

Feasibility study

Feasibility study in the sense it's a practical approach of implementing the proposed model of system . Here for a machine learning projects .we generally collect the input from online websites and filter the input data and visualize them in graphical format and then the data is divided for training and testing . That training is testing data is given to the algorithms to predict the data .

1. First, we take dataset.

2. Filter dataset according to requirements and create a new dataset which has attribute according to analysis to be done

3. Perform Pre-Processing on the dataset

4. Split the data into training and testing

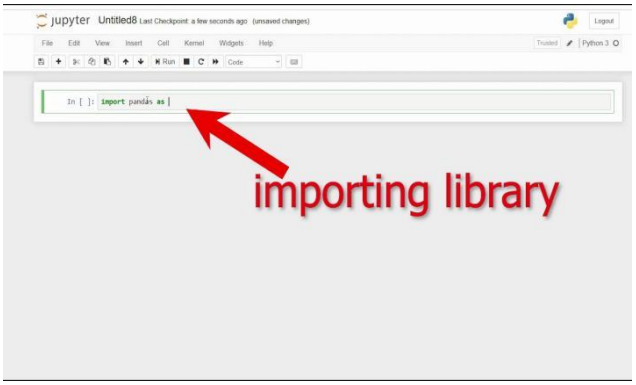
5. Train the model with training data then analyze testing dataset over classification algorithm

6. Finally you will get results as accuracy metrics.

RESULTS

Screenshots

i. **Importing packages**



ii. **Data Collection**

```
[2]: df = pd.read_csv('file.csv')

[3]: df
```

[3]:

	Name	Score
0	a	90
1	b	80
2	c	95
3	d	20

iii. **Data preprocessing**

```
In [155]: #drop the records with age missing in inp0 and copy in inp1 dataframe.
          inp0['age'].dropna(inplace=True)

In [156]: inp0['age'].isnull().sum()
Out[156]: 0

In [157]: inp0.isnull().sum()
Out[157]: age          20
          salary        0
          balance       0
          marital       0
          targeted      0
          default       0
```

iv. **Feature Extraction**

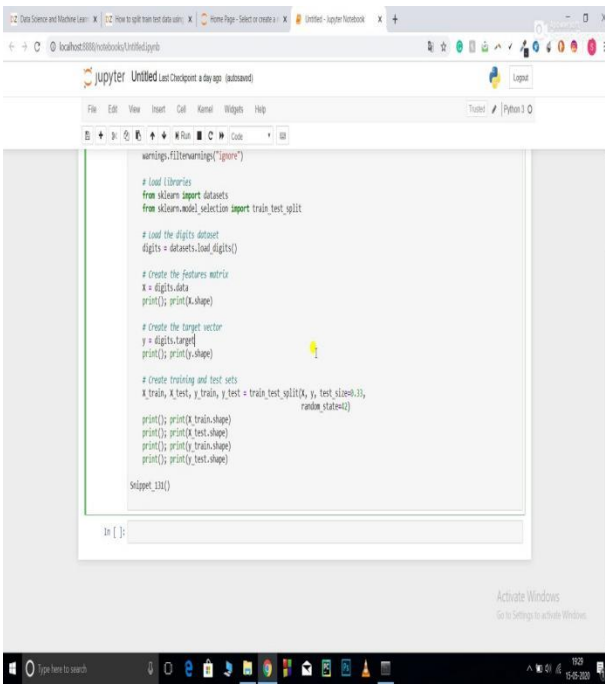
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijsem.org

Vol 14, Issuse.2 April 2023

v. **Training and Testing**



vi. **Evaluation model**

```
from sklearn.ensemble import RandomForestRegressor
ran_for = RandomForestRegressor()
ran_for.fit(X_train, Y_train)
Y_pred_ran_for = ran_for.predict(X_test)
r2=r2_score(Y_test, Y_pred_ran_for)
print("R2 score:", r2_score(Y_test, Y_pred_ran_for))
```

```
R2 score: 0.7226846570984489
```

RESULTS

In this section, we address the outcomes of this review. This subsection gives an overview of the selected papers of this review. The results of each research question are addressed in detail in the following five sections. A total of 290 stud- ies were chosen which implemented machine learning for anomaly detection.

CHAPTER 8: CONCLUSION

8.1 Conclusion

Anomaly detection techniques are mainly divided into two classifications: machine learning based, and non-machine learning based. The non-machine learning based techniques can be classified into statistical and knowledge based. Regarding this review, there are 274 articles that discuss the detection of anomalies through machine learning tech- niques. On the other hand, there are 16 articles that focus on non-machine learning based techniques. Detection of anomalies can be used in a wide variety of applications. In this review, we identified 43 different appli- cations in the selected papers.

CHAPTER 9: FUTURE ENHANCEMENTS

9.1 Future enhancements

Disruptions happen due to changes and they may cause severe outages impacting your business. But if you can see the changes that are happening in your environment in near real-time, you can prevent these disruptions, and thus the outages. In today's digital business environment where a significant portion of the business runs through applications, being reactive to outages isn't an option anymore.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, ''Anomaly detection: A sur- vey,'' ACM Comput. Surv., vol. 41, no. 3, pp. 71–97, 2009, doi: 10.1145/1541880.1541882.

[2] M. Injadat, F. Salo, A. B. Nassif, A. Essex, and A. Shami, ''Bayesian optimization with machine learning algorithms towards anomaly detec- tion,'' in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2018, pp. 1–6, doi: 10.1109/GLOCOM.2018.8647714.

[3] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, Unsupervised Anomaly Detection With Generative Adversar- ial Networks to Guide Marker Discovery, vol. 10265, no. 2. Cham, Switzerland: Springer, 2017.

[4] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, ''Data mining techniques in intrusion detection systems: A systematic lit- erature review,'' IEEE Access, vol. 6, pp. 56046–56058, 2018, doi: 10.1109/ACCESS.2018.2872784.

[11] Nov. 2012.

[5] F. Salo, M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Essex, ''Clustering enabled classification using ensemble feature selection for intrusion detection,'' in Proc. Int. Conf. Comput., Netw. Commun. (ICNC), 2019, pp. 276–281, doi: 10.1109/ICCNC.2019.8685636.

[6] F. Salo, A. B. Nassif, and A. Essex, ''Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection,'' Comput. Netw., vol. 148, pp. 164–175, Jan. 2019, doi: 10.1016/J.COMNET.2018.11.010.

[7] P. Gogoi, D. K. Bhattacharyya, B. Borah, and J. K. Kalita, ''A survey of outlier detection methods in network anomaly identification,'' Comput. J., vol. 54, no. 4, pp. 570–588, Apr. 2011, doi: 10.1093/comjnl/bxr026.

[8] S. Agrawal and J. Agrawal, ''Survey on anomaly detection using data mining techniques,'' Procedia Comput. Sci., vol. 60, no. 1, pp. 708–713, 2015, doi: 10.1016/j.procs.2015.08.220.

[9] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, ''Real-time big data processing for anomaly detection: A survey,'' Int. J. Inf. Manage., vol. 45, pp. 289–307, Apr. 2019, doi: 10.1016/j.ijinfomgt.2018.08.006.

[10] V. Chandola, A. Banerjee, and V. Kumar, ''Anomaly detection for discrete sequences: A survey,'' IEEE Trans. Knowl. Data Eng., vol. 24, no. 5, pp. 823–839,