Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 8. No. 1, 2020

Article Citation Format
1Ukoba J.O., 2Ochei C.L & 3Okengwu U.A. (2020): A Review on Anomaly Detection and Classification in Students' Assessment Dataset. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 8, No. 1. Pp 89-96

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received: 13th December, 2019
Review Type: Blind
Final Acceptance: 11th March, 2020

A Review on Anomaly Detection and Classification in Students' Assessment Dataset

[1]Ukoba J.O., [2]Ochei C.L & [3]Okengwu U.A.
Department of Computer Science
University of Port Harcourt
Nigeria.
[1]E-mail: [1]ukobajosephorode@gmail.com

ABSTRACT

Anomaly detection is a problem that has been of great interest in diverse research fields, which is aimed at finding patterns or behaviours that do not conform to the expected one. Many anomaly detection techniques has be applied in detecting anomalies in various fields, such as in intrusion detection, computer networks, health etc. In this review paper, we looked at an overview of anomaly detection techniques and detecting anomalies in students' assessment dataset. Thereafter we classified the reviewed papers/articles on anomaly detection in students' assessment data and the use of SVM and K-NN classifiers in detecting anomalies. It was observed that very little has been done with regards detecting anomalies in students' results/assessment scores. SVM is very good in detecting anomalies in datasets with high data instances and features with high sparsity. Most interesting areas of applications as reviewed in this paper are in intrusion detection, healthcare (medicine) and computer networks.

Keywords: Anomaly Detection, Student Assessment dataset, Normal distribution and K-Nearest Neighbour (K-NN) Classification

1. NTRODUCTION

Analysis of anomalies is of great interest of diverse research fields, including data mining and the machine learning approach (Xiaodan Xu et al, 2019). Behaviors or patterns in a data set that do not conform to a well defined normal behavior are referred to as anomalies, which may be caused by a malicious activity or some kind of intrusion (Agrawal, 2015). Anomalies are also referred to as outliers, surprise, aberrant, deviation, peculiarity etc (Chandola et al, 2009). This abnormal behavior found in the dataset is interesting to the analyst and this is the most important feature for anomaly detection (Dokas et al, 2002). Identifying these unexpected behaviors or patterns is what anomaly analysis is aimed at (Xiaodan Xu et al, 2019). Anomaly detection is a problem of finding patterns that do not conform to the expected or existing behavior (Chandola et al, 2009). It refers to detecting and identifying patterns in a given data set that do not conform to an established, normal behavior (Shahreza et al 2011). It has many applications in varieties of domains and fields (Shahreza et al 2011).

In the medical sciences, it is applied in the health monitoring systems for spotting malignant tumors in an MRI scan and in computer networks as an intrusion detection system in identifying strange patterns in network traffic that could signal a hack. It has its uses in the banking institution to detect frauds in credit card transactions and also in fault detection in operating environments (Pramit; 2017). It is applied in cyber security for intrusion detection, in space craft sensor in detecting a fault, to detect anomalous traffic pattern in a computer network which could signify that a hacked computer is sending out sensitive information or data (Chandola et al, 2009, Kumar, 2005).This paper seeks to review and classify the application of anomaly detection in students' assessment scores using two supervised machine learning techniques: Support Vector Machine and K-Nearest Neighbor.

The rest of this paper is organized as follows: Section II gives an overview of anomaly detection techniques, and then an x-ray of anomaly detection in students' assessment dataset was carried out in Section III. In Section IV, we reviewed related work on students' assessment dataset and the use of SVM and K-NN in anomaly detection. The reviewed work is then classified based on students' assessment dataset and the approaches used in Section V. Section VI and VII covers the evaluation of the classification and discussion of findings. Then the conclusion and recommendation for future work is given in Section VIII and IX.

2. OVERVIEW OF ANOMALY DETECTION TECHNIQUES

Anomaly detection is a problem or process of fining, identifying or detecting anomalies. Anomalies are data points that are inconsistent with the distribution of the majority of data points (Chandola et al, 2009). Figure 1 illustrates anomalies in a simple 2-dimensional data set. The data has two normal regions, $\mathbf{N_A}$ and $\mathbf{N_B}$, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., point $\mathbf{O_A}$ and $\mathbf{O_B}$, and points in region $\mathbf{O_C}$, are anomalies (Chandola et al, 2009.

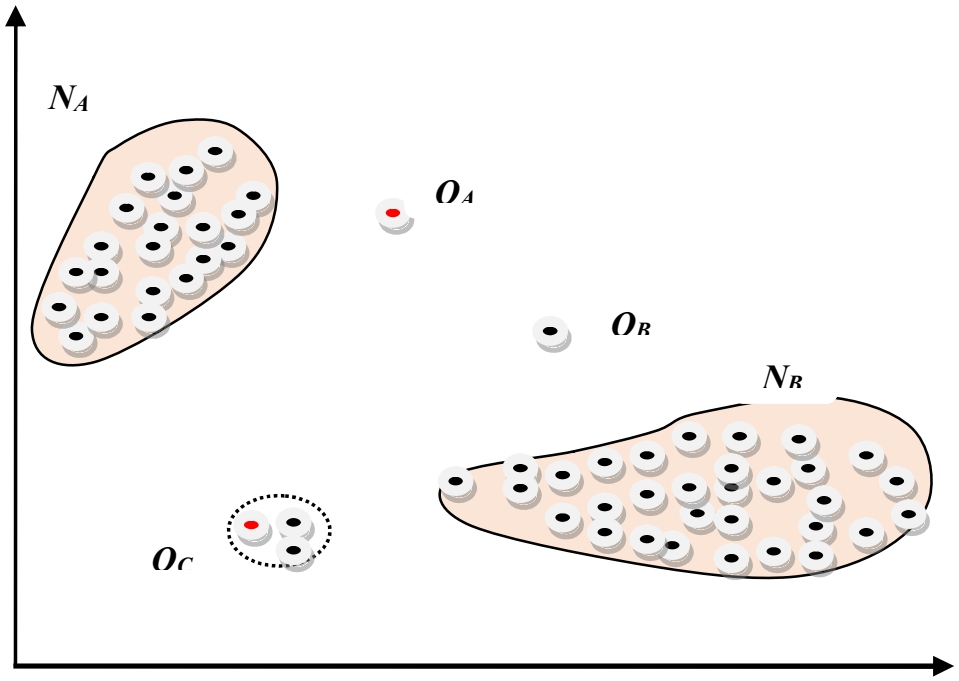

Figure 1: A simple example of anomalies in a two dimensional dataset

Anomalies can broadly be categorized into: point anomaly, contextual anomaly and collective anomaly (Chandola et al, 2009).

4. **Point anomaly:** A point anomaly is a single instance of data in a dataset that is too far from the rest. For example detecting credit card fraud based on amount spent (Chandola et al, 2009)..
5. **Contextual anomalies:** If a data point is anomalous in a specific context, but not otherwise, then it is termed a contextual anomaly which also known as conditional anomaly. This type of abnormality is context specific and is common in time-series data (Pramit, 2017). For example, high expenses during a non-festive month can be deemed a contextual anomaly (Ahmed et al, 2015).
6. **Collective anomalies:** This is an anomaly where a collection of related data instances is anomalous with respect to the entire data set. The individual data instances in a collective anomaly may not be anomalies by themselves, but their existence together as a collection is anomalous (Huysmans et al, 2005). For example, someone is trying to copy data form a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber attack.

An anomaly detection approach usually consists of two phases: a training phase and a testing phase. In the testing phase, the normal traffic profile is defined. In the training phase, the learned profile is applied to new data (Patcha and Park, 2007). Machine learning which focuses on classification and prediction based on known properties previously learned from the training data, is an aspect of artificial intelligence that gives computer the ability to learn without being explicitly programmed. Its major focus is to extract information from data automatically, by computational and statistical methods and thus closely related to data mining and statistics (Buczak and Guven, 2016; Svensson and Söderberg, 2008). Machine learning algorithms, unlike data mining need a goal (problem formulation) from the domain (Buczak and Guven, 2016).

Based on the context to which data labels are available, anomaly detection can be done using the following three machine learning techniques:

4. **Unsupervised Anomaly Machine Learning:** Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. (Brownlee, 2016). Examples are clustering and association problems.
5. **Semi-Supervised Machine Learning:** Semi-supervised learning is where you have a large amount of data **(X)** and only some of the data is labelled **(Y).** Semi-supervised learning sits in between supervised and unsupervised learning (Brownlee, 2016).
6. **Supervised Machine Learning:** Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. **Y = f(X)**. Examples are classification and regression problems (Brownlee, 2016). Common supervised machine learning algorithms are the K-Nearest Neighbourhood (KNN), Decision Trees (DT), Support Vector Machines (SVM), Bayesian etc.

3. DETECTING ANOMALIES IN STUDENT ASSESSMENT DATA

Assessing people's performance be it employees, representatives or students are one of the most important activities that take place in every sector and institutions. In companies and industries, overall performance of the staffs or employee are assessed quarterly or annually so as to provide necessary feed-back for effective management decisions. This involves the supervisors and management grading/assessing each staff according to his/her performance with respect to some metrics' for evaluation. This is then sent to the management board for decision making. Assessing the companies input and output is also a necessity so as to evaluate its performance and take necessary decision.

In education, assessments are by administering tests, assignments and examinations. The tests and assignments may constitute the continuous assessment. Examination is seen as the climax of every teaching and learning process (Salami et al, 2016). Students would have to take series of internal and sometimes external examinations before they could advance to the next stage. Assessing students' performance is key, as it evaluates the students' rate of assimilation and gives feedback for improving students learning. After teaching, examinations are set for the students. It is only when the student passes the assessment (examination) above the cut of mark that he/she can advance further to another level. These cut of mark is agreed upon by the management based on the general performance.

After examinations are written, course lecturers assess the student answers and allocate marks to the student. The total score of the student which determines the grade of the student is computed by adding the continuous assessment (CA) score to the examination score. It is this score that is used to determine the performance of the students in any course. The collection of student's grade is used to compute the CGPA of the student, which in turn determines the class the student graduates with (Salami et al, 2016). An accurate result computation, compilation and approval bring the semester or term to end. Evaluating and approving students results manually is faced with some setback such as fatigue, waste of energy and time and the board/senate members inability to detect certain anomalies, because the large information they are to assimilate (Salami et al, 2016).

4. REVIEW OF RELATED WORK

Many researchers in recent times have applied anomaly detection in several domains and have been the topic of a number of surveys and review articles. Yoo and Kim (2014) applied anomaly detection techniques in identifying and detecting malicious web pages using a hierarchical framework of the C4.5 Decision Tree and One-class SVM. The C4.5 Decision Tree was used to detect known malicious pages. It then uses an anomaly detection model built on the one-class support vector machine to detect unknown malicious activities. This framework achieved a significantly high malicious web page detection rate. However, it has a slightly high false positive rate. Viswanath et al (2014) used Principal Component Analysis (PCA) for distinguishing potentially bad unwanted behavior from normal behavior in a social network. The PCA is an unsupervised anomaly detection technique which is used to find patterns in high dimensional data. After detection, the K Nearest Neighbor algorithm was used to identify the category of anomalous users. Experimental results show 66% detection rate with less than 0.3% false positive rate.

Salami et al (2016) focused on detecting anomalies in students' result using the Decision Trees. The decision tree model was able to detect efficiently anomalies in student results in most cases. However it was unable to detect or identify anomalies in a situation where the training dataset has few anomalous instances. Serkani et al (2018) developed a hybrid a hybrid intrusion detection (anomaly detection) system using the C5.0 Decision Tree to reduce features in the traffic data and LS-SVM for training the IDS.

They considered datasets containing 10 classes including one normal and 9 network attack types which are; analysis, Backdoor, DoS, Exploit, Generic, Reconnaissance, Shell code and Worm. Experimental result showed very high detection accuracy in all the nine classes. However, there was relatively high false positive rate in the Worm class.

Cortez and Silva (2008), examined the performance of decision trees, Random forest, Neural Network and Support Vector Machines on secondary school grades using binary classification, regression and five-level classification for evaluation. They explained that social, demographic and school related variables also affect students' performance. Xianbo et al (2018) proposed a SVM-BGPAD model to detect BGP update messages explosion anomalies using the SVM classification model. Result showed a very high detection rate in terms of accuracy at 91.36% and F1-score. However, when the difference between the anomaly and regular dataset is high, the accuracy and F!-score of the model is very poor.

Foitre et al (2018) compared the performance of SVM with other algorithms to detect anomalies in water quality. Result showed that SVM detected anomalies at a very high rate. However, the false negative rate is at 30% and thus still vulnerable. Kumar et al (2010) used the SVM classification algorithm to identify positive and negative results of heart tumor. Divya and Kumaran (2016) examined the performance pf the Classic K- Nearest Neighbor algorithm on patient dataset.

5. CLASSIFICATION OF ANOMALY DETECTION IN STUDENTS' ASSESSMENT DATA

The table below shows the Classification based on students' assessment data (Datasets)

Table 8: Classification based on students' assessment data (Datasets)

Author and Date	Problem tackled / Contribution	Techniques Used	Metrics used for evaluation	Challenges / Research Gaps
Salami et al (2016)	Detected course based and student based anomalies in students' result	Decision Trees	• Accuracy • Specificity • Sensitivity	Unable to detect anomalies where the training datasets has few anomalous instances
Cortez and Silva (2008)	Achieved high predictive accuracy of secondary school grades. Explanatory analysis shows social, demographic and school related variables also affects students performance	Decision Trees, \Random Forest, Neural Networks and Support Vector Machines	• Binary classification • Five-level classification • Regression	Learning was offline and thus contains little features of students' data.

Table 9: Classification bases on approach (Support Vector Machines and K-Nearest Neighbour)

Author and Date	Area of Application	Problem Tackled / Contributions	Type of dataset used	Techniques used	Metrics for Evaluation	Challenges / Research Gap
Xianbo et al (2018)	Computer Networks	Detected anomalies in Border Gateway Protocols (BGP) at very high accuracy and F1-Score	BGP dataset	SVM	➢ Accuracy ➢ F1-Score	Not suitable for detecting anomalies with low anomalous data compared to the regular data
Kumar et al (2010)	Medicine	Effectively identified positive and negative results of heart tumor	Healthcare (Patient) dataset	SVM	➢ Sensitivity ➢ Specificity	
Yoo and Kim (2014)	Cyber Security	Identified and detected malicious web pages at a very high rate	Webpage dataset	A frame work of C4.5 DT and one-class SVM	➢ Accuracy ➢	Produced a slightly high false positive rate
Serkani et al (2018)	Intrusion Detection	Detected possible intrusions at a very high accuracy rate	Wired network dataset	Hybrid of C5.0 DT and LS-SVM	➢ Accuracy ➢ Gain Ratio	False positive rate for worm attack was relatively high
Tsigkritis et al (2018)	Intrusion detection on	Detected threatening behaviour in communication networks or cloud at a very high rate	Communication network dataset	K-NN	➢ Precision ➢ Recall	Requires scaling and normalization
Divya and Kumaran (2016)	Medicine (Health)	Yields superior data utility with a short time performance	Healthcare (Patient) datasets	Classic KNN	➢ Accuracy	Incurs high computational overhead
Foitre et al (2019)	Water Quality	Very high detection rate and low false positive rate.	Water quality time series datasets	SVM	➢ F-Score	False negative rate is at 30% and thus, still vulnerable

5.1 Evaluation of Classification

Table 1 shows the classification of the reviewed papers on anomaly detection based on students' assessment datasets. Table 2 shows the classification of the reviewed papers based on the approach used in terms of Support Vector Machines and the K-Nearest Neighbor algorithm.

5.2 Discussion of Findings

From Table 1, it is observed that little research has been done on detecting anomalies in students' assessment datasets. It also shows that the classification algorithms are mostly used in detecting anomalies in students' assessment dataset. This may be because, the students assessment dataset can be labelled easily as to which is normal or abnormal. Hence, we infer that the supervised machine learning algorithms is the most appropriate and easiest machine learning algorithm in detecting anomalies in students' assessment dataset.

Table 2, shows that SVM is most suitable in analysing and classifying anomalies when applying anomaly detection in health care, intrusion detection and computer networks. This is because most features in the health care dataset have zero value. SVM is also the best classifier for document classification problems where sparsity is high and the feature/instances are also very high which are characterized by healthcare, intrusion detection and computer network datasets. The K-NN on the other hand is used when there are many instances (points) an few dimensions or data.

6. CONCLUSION

There is no best algorithm for detecting and classifying anomalies. Each algorithm has its strengths with regards its application domain and nature of dataset used. It is very easy to label students' assessment dataset on which is normal or anomalous. Detecting anomalies in students' assessment score is a classification problem with many data instances and features.

7. RECOMMENDATION FOR FURTHER RESEARCH

We recommend that researches on detecting and classifying anomalies should spread to other domains especially the education sector with regards to assessment and examination scores. This would help make sure that results issued to students at different levels and at interviews reflects their actual and true performance.

REFERENCES

1. Agrawal S. and J. Agrawal (2015). Survey on Anomaly Detection Using Data Mining Techniques, Procedia Computer Science: 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 6, 708 - 713
2. Ahmed M., A. N. Mahmood, Jiakun (2015). A survey of network anomaly detection and technigues, Journal of Network and Computer Applications, 60, 19-31.
3. Brownlee J. (2016). Supervised and unsupervised machine learning algorithms. Retrieved on 20th June, 2019 from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
4. Buczak A. L and E. Guven (2016). A survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, IEE Communication Surveys, 24, 1153 – 1176
5. Chandola V., A. Banerjee and V. Kumar (2009). Anomaly Detection: A survey, ACM Computing, 59, 1-57.
6. Cortez P. and A. Silva (2008). Using data mining to predict secondary school students performance. EUROSIS..
7. Dokas P., L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, P.N Tan (2002).Data mining for network intrusion detection, In Proceedings of NSF Workshop on Next Generation Data Mining; 10, 21-30
8. Fitore M. , D. Logofatu, and F. Leon (2018). Approaches to building a detection model for quality water, Journal of Information and Communication, 23, 41-63.
9. Fitore M. , D. Logofatu, and F. Leon (2018). Machine learning approaches for anomaly detection of water quality on a real-world dataset, Journal of Information and Communication, 23, 41-63.
10. Hamza O. S., S. I. Ruquyyah, M. O. Yhaya (2016). Detecting Anomalies in Students' Results Using Decision Trees, International Journal of Modern Education and Computer Science, 7, 31-40
11. Huysmans J., B. Baesens, D. Martens, K. Denys and J. Vanthienen (2005). New Trends in Data Mining, Tijdschrift voor Economie en Management, 15, 1-14
12. Kumar V. (2005). Parallel and Distributed Computing for Cyber Security Distributed Systems Online. IEEE 6, 10
13. Parmar J. D. and J. T, Patel (2017). Anomaly Detection in Data Mining: A Review. International Journal of Advanced Research in Computer Science and Software Engineering, 9, 32- 40
14. Patcha A. and J. Park (2007). An overview of anomaly detection techniques: Existing Solutions and latest technological trends, scienceDirect, Elesevier, www.elsevier.com/locate/comnet.
15. Pramit C. (2017). Introduction to Anomaly Detection, Data Science.com. Retrieved on 4th April, 2019 from http/datascience.com/blog/python-anomaly-detection.html.
16. Serkani E., H. G. Garakani, N. M. Zadeh, E. Vaezpour (2018). Hybrid anomaly detection using decision tree and SVM, International Journal of Electrical and Computing Engineering, 11, 19-31.
17. Shahreza L. M. , D. Moazzami, B. Moshiri, M.R. Delavar (2011). Anomaly Detection using a Self Organizing Map and Particle Swarm Production, Scientia Iranica, 18(6), 1460 - 1468
18. Svensson M. and J. Söderberg 2008). Machine learning technologies in telecommunications, Ericsson Review, 4, 29-33
19. Viswanath B., M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy and A. Mislove (2014). Towards detecting anomalous user behaviour in online social network, Proceedings of 23rd USENIX Security Symposium. San Diego, CA: USENIX, USENIX Security Symposium (USENIX Security 14), 16, 223-238.
20. Xianbo D., N. Wang and W, Wang (2019). Application of machine learning in BGP anomaly detection. Journal of Physics: Conference Series. 1176(3), 10, 1742-6596.
21. Xiaodan X., H. Liu and M. Yao (2019). Recent Progress of Anomaliy detection. Complexity, 11, 2686378
22. Yoo S. and S. Kim (2014). Two-phase malicious webpage detection scheme using misuse and anomaly detection, International Journal of Reliable Information and Assurance, 2(1), 1 – 9