A Brief Introduction to Anomaly Detection and its Techniques

Tatwadarshi P. Nagarhalli
Computer Engineering Department
VIVA Institute of Technology
University of Mumbai
Mumbai, India
tatwadarshipn@gmail.com

Ashwini M. Save
Computer Engineering Department
VIVA Institute of Technology
University of Mumbai
Mumbai, India
ashwini.save@gmail.com

Abstract **-** Anomaly detection is subfield of machine learning where a model is developed which looks out for any abnormalities in the data. In the recent times of internet connectivity anomaly detection plays a very important role, and is a very important force multiplier. Also, anomaly detection has a very wide range applications in various domain. Due to these reasons extensive research has been carried out in this field. This paper introduces, as to, what does anomaly detection actually mean and also its techniques. The paper introduces the different application domains in which anomaly detection plays a very critical role.

Index Terms – *Anomaly Detection, Artificial Intelligence, Machine Learning, Fraud Detection.*

*****

I. INTRODUCTION

Anomaly Detection is a way or a process or a problem of finding patterns in a dataset whose behaviour is not as expected or do not conform to the expected behaviour [1]. The reason for extensive research in the field of anomaly detection is its wide ranging applications, like for example in fraud detection for credit or debit cards, insurance or health care, intrusion detection for cyber-security, fault detection in safety critical systems, military surveillance for enemy activities or identity thefts. This wide ranging application mandates a detail study of anomaly detection and its techniques.

II. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Anomaly detection is a subfield or a specialisation of Machine Learning, which itself is a subfield of Artificial Intelligence. So, before jumping into anomalies and anomaly detection it would be prudent to understand what is artificial intelligence and machine learning.

The field of Artificial Intelligence attempts to understand and build intelligent entities. Over a period of time many definitions have been proposed by many scholars but fundamentally the field artificial intelligence is the development of computer systems able to perform tasks normally requiring human intelligence. These tasks can be anything, like for example speech recognition or processing of natural language or decision making, etc [2].

Machine Learning is a further application or an extension of artificial intelligence. Now, for a computer to solve a problem it required an algorithm. But for some tasks we do not have algorithms, like for example detecting spam mails to a genuine one. For such type of problem where there is no algorithms to find a solution we try to learn from the data and try to solve the problem, that is what we lack in knowledge we try to cover it with data [3].

In machine learning we provide the machine learning algorithms with the large number of dataset to learning from; or in other words the machine learning algorithm trains itself with the help of the training set or the data set provided to it. These algorithms review these dataset and try to find some kind of relationships among them. So that for any unknown data set it can predict outcome with the help of these relationships that the machine learning algorithm has found out from the training set.

Over the years many machine learning algorithms have been proposed by different scholars. All these different machine learning algorithms fall into three major categories. These categories are Supervised Learning, Unsupervised Learning and Reinforcement Learning.

If in the training set all the different inputs have been provided with the expected outputs then this type of training set is called as labelled training set or labelled training data. If on the other hand for the different inputs expected outputs have not been provided then it is called as unlabelled training data [4]. In Supervised Learning the training algorithm is provided with a labelled training set [5], whereas if the training dataset is an unlabelled training set then it is called as Unsupervised Learning [6]. On the other hand the reinforcement Learning is somewhere between supervised and unsupervised

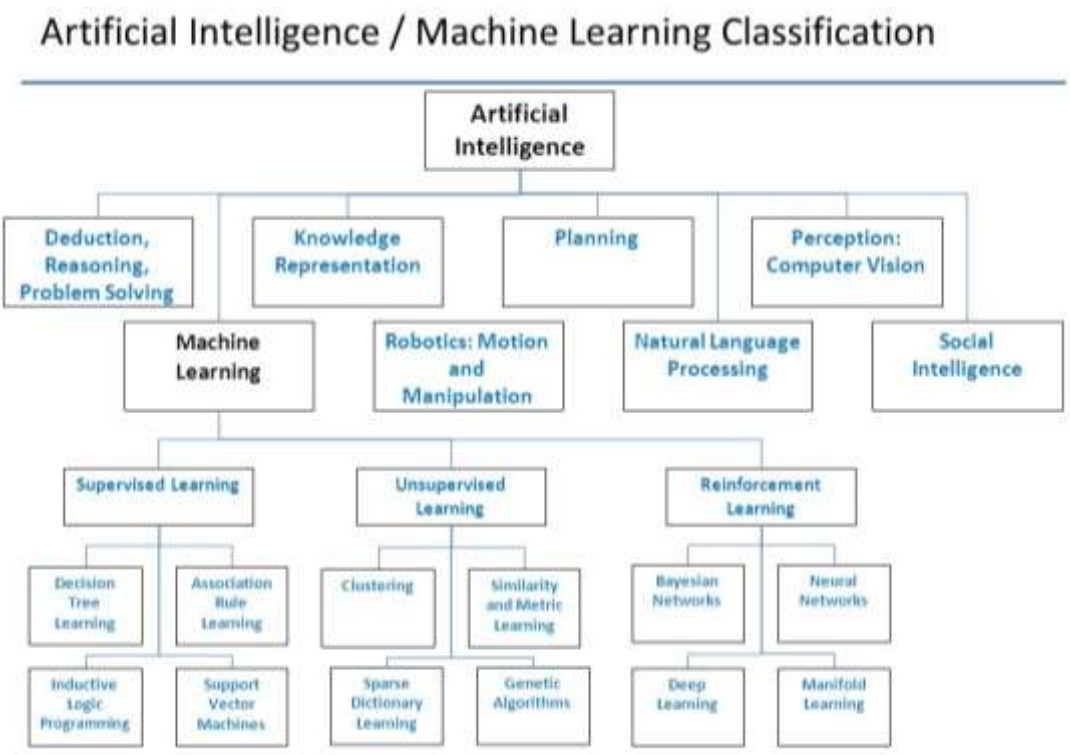

Fig. 1 Classification of Artificial Intelligence and Machine Learning [8]

learning. The training set provided to the learning algorithm is unlabelled, but when the algorithm predicts a wrong value then it is told that it has wrong [7].

Fig. 1 shows the classification of Artificial Intelligence and Machine Learning.

III. ANOMALY DETECTION

Anomaly detection can broadly classified into three categories. Anomaly detection with supervised learning, anomaly detection with unsupervised learning and anomaly detection with semi-supervised learning [1].

In Supervised anomaly detection, the training set used is a labelled training set with well defined as "normal" and "abnormal" training data. Whereas in Unsupervised anomaly detection algorithms use unlabelled training and testing test dataset under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. And, Semi-supervised anomaly detection techniques construct a model representing normal behaviour from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

IV. ANOMALIES

Anomalies are data points or data patterns which do not conform to the prescribed patterns attained by the majority of datasets.

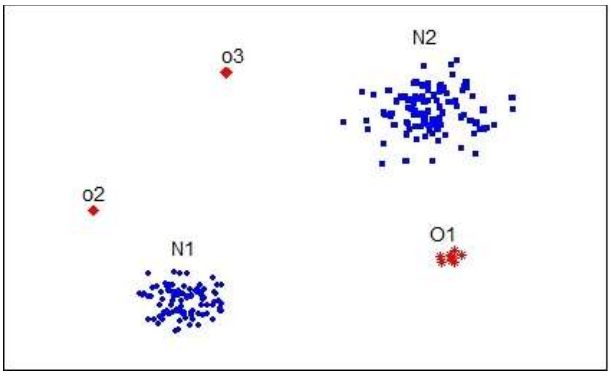

Fig.2 Outlier Example [9]

Take of example Fig. 2 which illustrates anomalies in a simple 2-dimentional hyperplane. The learning algorithm will take in the training set and will plot the data sets in a hyperplane as shown in Fig. 2. From the figure it can be seen that the region N1 and N2 form a sizeable region with many datasets conforming to it. Whereas, O1, O2 and O3 are those data points which do not come even

close to the two regions. These three data points or set of data points are known as anomalies or outliers.

V. ANOMALY DETECTION METHODOLOGY

There are many different ways in which anomaly detection can be conducted. But the basic methodology or the general outline of the anomaly detection remains the same. It can be summed up as given in the Fig. 3 [10].

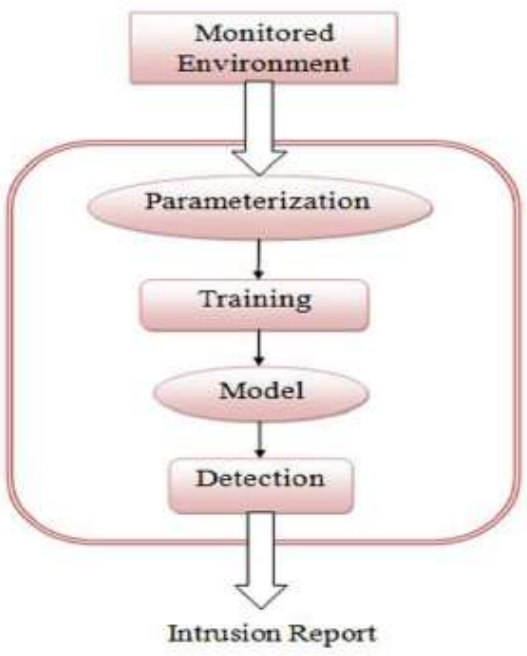

Fig. 3 Anomaly Detection Methodology

The Fig. 3 gives the general methodology for anomaly detection. In Fig. 3 the monitored environment is the different application environment where anomaly detection can be used, like for example intrusion detection, fraud detection, etc.

Parameterization means the preprocessing of the data. Preprocessing means working or converting the training data and the input data such that the learning algorithms can understand the data and learn from it or give proper prediction respectively. Without proper preprocessing the learning algorithm or the anomaly detection system will not be able to detect any anomalies accurately.

In training the anomaly detection system or model has to be trained properly. This training can be supervised, unsupervised or semi-supervised. The type of learning to be used will depend on the type of prediction model you are trying to build and the type of data that is available at your disposal.

In the detection step once the detection model has been trained appropriately the system will be able to predict any type of anomalies. Generally an alarm is triggered if the deviation is found to be over the set threshold.

For training and testing a learning algorithm for outlier detection there are many sources available online, including ODDS [11]. ODDS has a large collection of publicly available outlier detection datasets with ground truth in different domains.

VI. ANOMALY DETECTION TECHNIQUES

The anomaly detection techniques can be broadly divided into six categories [1].These are Classification based anomaly detection, Nearest Neighbour based anomaly detection, Clustering based anomaly detection, Statistical anomaly detection, Information Theoretic anomaly detection, Spectral anomaly detection.

In Classification based anomaly detection the model is trained with a labelled training dataset which would contain the ‘normal’ and ‘abnormal’ data points. In the Nearest Neighbour anomaly detection the input dataset consists of the k closest training examples in the feature space.In Clustering anomaly detection the dataset are having similar numerical values are grouped together. These groups are called as clusters. The data points not belonging to any of the clusters are the anomalous data points.

In Statistical anomaly detection some statistical means like probability are used to define the data points to be either anomalous or non-anomalous. Information Theoretic anomaly detection techniques use a concept of information entropy. It measures the uncertainty (or impurity) of a collection of data items.Spectral anomaly detection techniques try to find an approximation of the data using a combinationof attributes that capture the bulk of variability in the data.

There many ways in which the different techniques for anomaly detection can be viewed from. But it is rather important that it is looked from the application domains perspective. There are generally three major and critical application domain where anomaly detection plays a very important role, these are Intrusion detection, Fraud detection and Medical and Public Health anomaly detection.

A. Intrusion Detection

Intrusion detection is the process of identifying malicious activity targeted at computing and networking resources [12]. On the other hand Intrusion Detection System (IDS) monitors the computing environment and networks resources to identify any specious events or activities.

Gustavo Nascimento and Miguel Correia [13] compare different models that can be used for detecting intrusion with the help of software. The paper demonstrates the usefulness of using anomaly detection techniques for intrusion detection. The papers test many different learning models like Hidden Markov Maodel, N-gram model, etc.

Whereas V. Jyothsna and V. V. Rama Prasad [14] perform study of different anomaly based intrusion detection systems. The paper identifies many different techniques for anomaly detection for intrusion detection including,

Statistical based, Operational or threshold metric model, Markov Process or Marker Model, Statistical Moments or mean and standard deviation model, Univariate Model, Multivariate Model, Time series Model, Cognition based, Finite State Machine Model, Description script Model, Adept System Model, Machine Learning based, Baysian Model, Genetic Algorithm model, Neural Network Model, Fuzzy Logic Model, Outlier Detection Model, Computer Immunology based, User Intention based.

P. Garcı´a-Teodoroaet. al. [15] conduct a review of the most well-known anomaly-based intrusion detection techniques and systems. The paper also outline the main challenges to be dealt with for the wide scale deployment of anomaly-based intrusion detectors. Some of the challenges for the wide anomaly based intrusion detection systems identified in the paper include, low detection efficiency, low throughput and high cost, the absence of appropriate metrics and assessment methodologies.

Ke Wang and Salvatore J. Stolfo[16] on the other hand propose a new technique for intrusion detection. The paper proposes a system called as the PAYL which is a pay load based anomaly detector. The PAYL models the normal application payload of network traffic in a fully automatic, unsupervised. The paper first computes a training phase a profile byte frequency distribution and their standard deviation of the application payload flowing to a single host and port. Then Mahalanobis distance is used during the detection phase to calculate the similarity of new data against the pre-computed profile. The detector compares this measure against a threshold and generates an alert when the distance of the new input exceeds this set threshold.

B. Fraud Detection

Malicious users may target commercial institutions like banks or credit card companies or loan offices. These malicious users might be employees or impersonators of customers. Over the years many techniques and ways have been proposed to tackle fraud [17].

There are different ways through which fraud detection can be accomplished. Ray-I Chang et. al. [18] and RaghavendraPatidaret. al. [19] use neural networks to tackle this problem. WhereasTao Guoet. al. [20] and [21] make use of genetic algorithm for the learning purpose. Anita B. Desai et. al. showcase how other data mining techniques can be applied to detect credit card frauds. On the other hand S. Rossetet. al. [23] demonstrate that even the basic of learning algorithm like the logistic regression can be effectively used to deal with.

Miklos A. Vasarhelyi and Hussein Issa [24] describe classification-based and clustering-based anomaly detection techniques and their applications. That is more specifically the application to the problem ofcertainfraudulent activities has been examined in this paper. As an illustration, the paper applies K-Means, a clustering-based algorithm, to a refund transactions dataset with the intent of identifying fraudulent refunds.

C. Medical and Public Health

Another important domain where accurate anomaly detection plays a very critical role and requires a very high degree of accuracy. In this domain the data is generally patient records. The data can have anomalies due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Several techniques have also focussed on detecting disease outbreaks in a specific area.

Many scholars have used many different machine learning techniques like Naïve Bayesian classifier, Neural network, Rule based classifier, etc. to different degrees of accuracy [1].

There are many other application domains as well where Anomaly Detection plays a very critical role like textual analysis, sensor data analysis, etc. And all most all the proposed techniques in these application domain use or propose anomaly detection technique which will fall into one of the different categories that has already been discussed.

VII. CONCLUSION

In this paper a brief introduction has been given for artificial intelligence, machine learning and finally anomaly detection and its different techniques has been discussed. The critical application domain has been also been discussed. Also, the categorisation of anomaly detection techniques has also been discussed. The different anomaly detection techniques has been focused around the critical application domain so as to stress the importance of anomaly detection in various field.

REFERENCES

[1] V.Chandola, A. Banerjee and V. Kumar, “Anomaly Detection: A Survey”, ACM Computing Surveys, September 2009.
[2] S. J. Russell and Peter Norvig, “Artificial Intelligence: A Modern Approach”, Prentice Hall, New Jersey, USA, 1995.
[3] E.Alpaydin, “Introduction to Machine Learning”, Second Edition, The MIT Press Cambridge, Massachusetts London, England, 2010.
[4] P. Harrington, “Machine Learning in Action”, Manning Publications Co., NY, USA, 2012.
[5] T. Mitchell, “Machine Learning”, McGraw-Hill Science/Engineering/Math publications, NY, USA, March 1, 1997.
[6] W. Hsieh, “Machine Learning Methods in Environmental Sciences”, Cambridge University Press, Delhi, 2009.
[7] S.Marsland, “Machine Learning an Algorithmic Perspective”, CRC Press, NY, USA, 2009.
[8] M. Y. Wang and C. E. Zwilling, “Multivariate Computing and Robust Estimating for Outlier and Novelty in Data and Imaging Sciences”, Open Access (http://www.intechopen.com/books/advances-in-bioengineering/multivariate-computing-and-robust-estimating-for-outlier-and-novelty-in-data-and-imaging-sciences), 2015.
[9] http://slideplayer.com/slide/7002258/, Last accessed on 25th January, 2017.

[10] S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Published by Elsevier B.V., 2015, pp. 708-713.

[11] http://odds.cs.stonybrook.edu/, last accessed on 25th January, 2017.

[12] E. G. Amoroso, "Intrusion Detection", Intrusion, Net Books, 1999.

[13] G.Nascimento and M.Correia, "Anomaly-based Intrusion Detection in Software as a Service", INESC-ID – Portugal, http://wraits11.di.fc.ul.pt/papers/nascimento-webintdetect.pdf.

[14] V. Jyothsna and V. V. Rama Prasad, "A Review of Anomaly based IntrusionDetection Systems", International Journal of Computer Applications, Volume 28– No.7, September 2011, pp. 26-35.

[15] P. Garcı´a-Teodoroa, J. Dı´az-Verdejoa, G. Macia´-Ferna´ndeza , E. Va´zquezb, "Anomaly-based network intrusion detection: Techniques, systems and challenges", Published by Elsevier, 2009, pp. 18-28.

[16] K. Wang and S. J. Stolfo, "Anomalous Payload-based Network Intrusion Detection", http://www.covert.io/research-papers/security/PAYL%20-%20Anomalous%20Payload-based%20Network%20Intrusion%20Detection.pdf

[17] L. N.Lata, I. A.Koushika and S. S. Hasan, "A Comprehensive Survey of Fraud Detection Techniques", International Journal of Applied Information Systems (IJAIS), 2015, pp. 26-32.

[18] R. Chang, L. Lai, W. Su, J. Wang and J.Kouh "Intrusion Detection by Backpropagation Neural Networks with Sample-Query and Attribute-Query", Research India Publications; (2006)., pp. 6-10.

[19] R.Patidar and L. Sharma "Credit Card Fraud Detection Using Neural Network". International Journal of Soft Computing and Engineering (IJSCE), 2011, Volume-1, Issue1, pp. 32-38.

[20] T. Guo and G. Li "Neural Data Mining For Credit Card Fraud Detection". IEEE, Proceedings of the Seventh International Conference on Machine Learning and Cybernetics; (2008). (3630-3634).

[21] Review Paper on Credit Card Fraud Detection, Suman Research Scholar, GJUS&T Hisar HCE Sonepat, NutanMtech. CSE, HCE Sonepat.

[22] A. B. Desai and Dr. R.Deshmukh, "Data mining techniques for Fraud Detection", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, pp. 1-4.

[23] S.Rosset, U. Murad, E. Neumann, Y. Idan, and G. Pinkas, "Discovery of fraud rules for telecommunications challenges and solutions", In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, 1999, pp. 409-413.

[24] M. A. Vasarhelyi and H.Issa, "Application of Anomaly Detection Techniques to Identify Fraudulent Refunds", SSRN Electronic Journal, August 2011, http://ssrn.com/abstract=1910468.