Survey of anomaly detection in high dimentional big data

JAYKUMAR SAVANI*,

TU-Dortmund,

Germany

Anomaly detection in high-dimensional data is rapidly becoming a basic research subject with a wide range of real-world applications. Many existing methods for anomaly detection fails to maintain acceptable accuracy for generated data that is very high in volume and velocity. Accuracy and performance achieved by traditional methods are affected by the curse of dimensionality problem, might be referred to as this occurrence of having both difficulties simultaneously. To close this gap and grasp the underlying issue, it's vital to identify the particular problems when dealing with both big data and higher dimensionality. As a result, the goal of this survey is to provide the overview of anomaly detection in context of higher dimensional big data by identifying the unique challenges including the problem of high dimensionality, techniques/algorithms (in anomaly detection) and tackling methods. Furthermore, the limits of old methodologies and contemporary high-dimensional data strategies are explored as well as provide an overview of comparison between dimensionality reduction methods.

Additional Key Words and Phrases: Big data analytics, High dimensional data,Curse of dimensionality, Outlier Detection, Anomaly Detection, Anomaly detection techniques, PCA, Auto-encoders

ACM Reference Format:

Jaykumar Savani. 2022. Survey of anomaly detection in high dimentional big data. 1, 1 (February 2022), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Health records, Linked data (TCP/HTTPS data logs), financial data, and as well as voice network data, business, , and biomedical or bioscience,etc are generated at a rapid pace and large scale [1, 2]. This subject has recently become a focus of academia, and is referred to as "big data," a term that represents the massive and spread nature of the data sets. According to Gartner [3], Big data is described as velocity, veracity and volume at larger scale data sets that required innovative and cost-effective data analysis to make decisions and acquire relevant information. In recent years, big data problems has been arises. In figure 1, the six V's of big data: , veracity, variety, volume, velocity, value and variability is shown [4][37]. Anomaly detection in huge data sets becomes more difficult as the number of dimensions, characteristics, or attributes grows.

[*] Matriculation Number:230443

Author's address: Jaykumar Savani, jaykumar.savani@tu-dortmund.de,
TU-Dortmund,
Dortmund, North Rhine-Westphalia,
Germany, 44227.

Manuscript submitted to ACM

The major goal of paper is to provide an overview of the current scenario of what are the specific challenges, techniques/algorithms (anomaly detection) of anomaly detection in high-dimensional big data , and techniques/algorithms to overcome the curse of dimensionality. In addition, the limits of conventional methodologies and contemporary tactics for high-dimensional data, as well as the most recent techniques and big data applications necessary for anomaly detection optimization, are also covered.

2 ANOMALY DETECTION IN CONTEXT OF BIG DATA

The majority of the applications were for large data sets with hundreds or even millions of characteristics. It is difficult to identify anomalies in higher dimension spaces due to higher sparsity and also arise in subset of subspaces[5]. Thus, distance based techniques suffer very often in higher dimension space and leads to the changes in traditional algorithms[6]. Furthermore, typical approaches lose their use as data dimensionality rises because they rely on tactics that make assumptions about the data's low dimensionality [7]. Furthermore, for data sets with high dimensionality, only a proportion of data points may be useful [6].

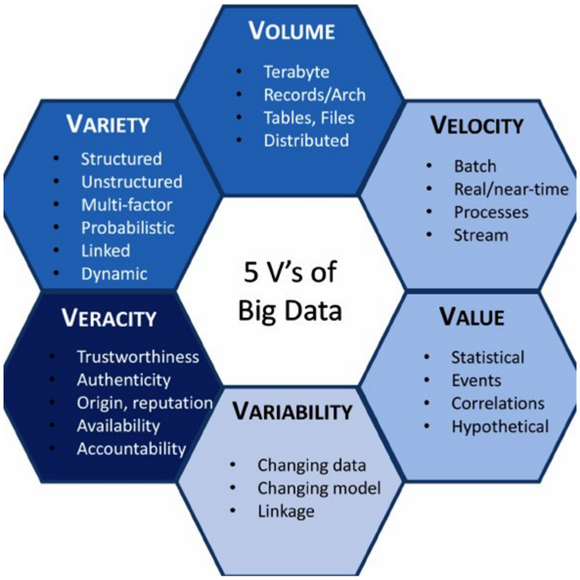

Fig. 1. Characteristics of big data

Anomaly detection may be used in both live streamed data and historically stored data to solve issues with high dimensionality. For historical data, batch processing techniques is used. Since the datasets are stored over time periods, it related to the bulk or higher volume size characteristic. In contrast, for live streamed data, newly generated data points or observations processed continuously through various streams known as data stream to identify the anomaly in real time (velocity aspect of big data). In many domains, such as machine learning and data mining, a number of existing studies and evaluations emphasize the challenge of high dimensionality. Challenges like uncertainty of data, performance of the technique (both memory and computational time), scalability arises in perspective of volume of data. Apart from that, asynchronous instances of data streams, dynamic relationships (For example: correlations between features), schema heterogeneity, etc needs to be faced from velocity perspective. The "size" aspect of volume

[8][9][10][11][12] and the "speed" aspect of velocity [13][14][15][16][17][18] are extensively discussed in the literature, but the "dimension" aspect is usually overlooked.

The theoretical foundation and existing methodologies are described in the following sections to determine the specific problems posed by anomaly detection in higher dimension and to comprehend the underlying difficulty impeding the accuracy and performance of traditional techniques.

3 THE CURSE OF DIMENSIONALITY PROBLEM

Bellman [21] used the phrase "curse of dimensionality" to characterize the issue produced by an increase in the number of input features or dimensions. Data sparsity results from an increase in data dimensionality, which causes an increase in data size correspondingly. In addition, the analysis of sparse data is also challenging.The curse of dimensionality has an impact on anomaly detection approaches and it makes more difficult to identify exact anomalous data point masked or hidden by unneeded features[5]. Organizations with large transaction volumes and accompanying databases are a major source of high-dimensional data. Combination of data processing and machine learning or deep learning methods manages the growing number of dimensions without compromising accuracy.

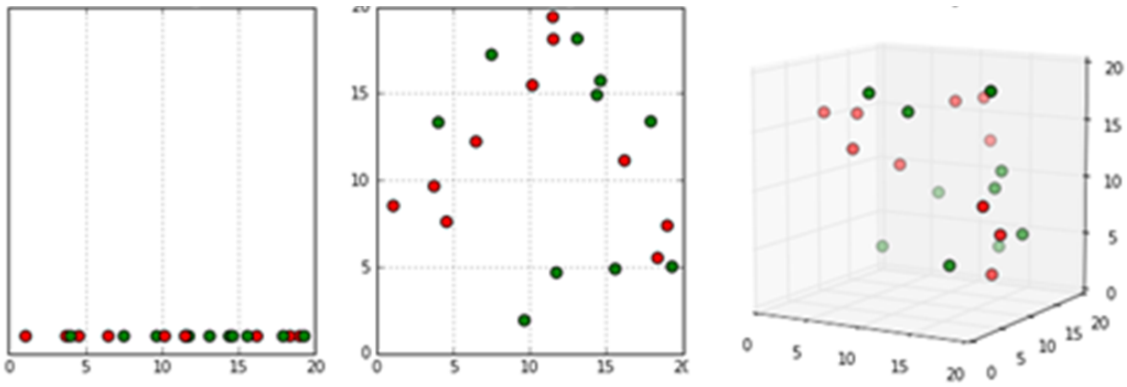

Fig. 2. Curse of dimesionality effects (a) 1-Dimension, (b) 2-Dimensions, (c) 3-Dimensions

Data points become fragmented and isolated as the number of dimensions increases, making it difficult to locate the data set's global optima. A data collection grows increasingly complicated as more dimensions are added, as each additional dimension introduces a significant number of false positives [22]. When projected in one, two, and three dimensions, Figure 2 depicts the data's sparsity [38]. It causes a slew of issues with various noise levels, such as inappropriate characteristics or needless attributes, which can muddle or even obscure data instances [7]. Many algorithms struggle with high-dimensional data because of such challenges. Due to the curse of dimensionality, statistical techniques such as distance measurements become less helpful as the number of dimensions grows, because the points become practically equidistant from each other. Many algorithms use proximity ideas to identify anomalies based on data point relationships. However, in high-dimensional space, such algorithms experience massive computational expansion and, as a result, lose their usefulness. According to Aggarwal [6], practically every approach that is largely focused on the idea of distance will suffer qualitatively in high-dimensional space, necessitating a more significant re-definition.Distance based approaches are much more simple for implementation wise (data point as a point of reference in spars dimensional space). No previous assumptions about the data distribution model are made, and locality may be defined in a variety of ways. This includes clustering, distance/proximity, density and classification based techniques. The overview of such techniques is discussed in following sections.

3.1 Cluster based techniques

Clustering based techniques are based on the similarity and heuristic difference among the observations. One of the efficient method is taken as an example in this section. Ertoz et al. [23] provide an idea of a shared nearest neighbor clustering technique. It is used to detecting clusters with varying densities, sizes, forms, and anomalies also. In this approach, number of clusters are automatically decide according to the shifting densities and multidimensionality. The approach consist of several steps. Initially, it detects each data instance's nearest neighbours. After that it recalculate the similarity between pairs of data instances in terms of the number of shared nearest neighbours. Then, Using similarity, the approach finds core points and builds clusters around them. With varied densities and rising dimensions, the shared closest neighbour concept of similarity reduces difficulties. The researchers discovered a number of enhancements that enable their method to analyse massive data sets effectively.[24]

3.2 Distance based techniques

Angiulli and Pizzuti [8] introduced a distance-based anomaly identification algorithm named "HilOut" to identify the outliers in highdimentional big data. To linearize the data set, HilOut employs the concept of the space-filling curve. The approximation solution is estimated in earlier phase. Then, by checking the remained outliers from first phase, the actual solution is calculated. They also provided comprehensive overview of this method in terms of scalability and different variants of HilOut such as disc based algorithm and in-memory algorithm. Instead of using similarity, dissimilarity between data points are much more efficient techniques to manage sparse data points in higher dimensions. To deal with the same problem, Koufakou and Georgiopoulos [25] also suggested a different method. Their method speedup the process to detect anomalies. It is obtained through distributed version that is very close to linear. It is also known as "fast distributed" methode and was designed for mixed-attribute data sets with sparse higher dimensional data. Since the data set contains numerous points and many characteristics [24], it takes into account the sparseness of the dataset and also higher scalability according to Koufakou and Georgiopoulos.

3.3 Density based techniques

The dense locales of the data space, characterized by diverse areas of lesser item density, are dealt with via density-based approaches. In higher dimension space, earlier density based techniques are also ineffective since the data points are dispersed and leads to fall in their density (due to curse of dimensionality). Thus, makes it more difficult to locate the relevant data points (or collection of data points)[26]. To overcome such challanges in density based techniques, Chen et al. [27] provides a density estimator method to estimate measures in high-dimensional data which they applied to the challenge of detecting changes in data distribution.[24]

3.4 Classifiaction based techniques

When it comes to classification and high dimensionality, a typical issue is when the feature vector's dimensionality m is substantially bigger than the available training sample size n. Method provided by Fan and Fan [28] describes excessive dimensionality in classification is inherent. According to them, it caused by irrelevant noise effects that obstruct the reduction of classification error. High dimensionality raises the certain challenge on classification accuracy, predictive power, and model interpretability. In another words, the ability of the model to estimate what kind of relationship exist between input data points and output results [28][24].

4 TECHNIQUES TO OVERCOME CURSE OF DIMENSIONALITY

In recent years there are various techniques has been developed to overcome the problem of curse of dimensionality challenge. One the method is called dimensionality reduction. This section describes the two of the dimensionality reduction method (PCA, Autoencoders)and comparison among these methods.

4.1 PCA (Principal component analysis)

PCA is a linear dimensionality reduction approach for extracting data from a high-dimensional space by projecting it onto a lower-dimensional sub-space.It seeks to keep the vital elements of the data that have the most variation and eliminate the non-essential sections that have the least variation[paper]. The major goal of PCA is to extract all of the important features from the data set and joined them into completely new orthogonal features known as primary components. In addition, it calculates the product of two distinct tiny matrices to estimate a data set, known as data matrix. The dimensionality is minimised through combination of all characteristics and identify required patterns of data from data matrix [29], referred to as primary components, before removing all other attributes [30,31]. The figure 3 gives an quick overview of the flow chart to perform a PCA on given set of data [32].

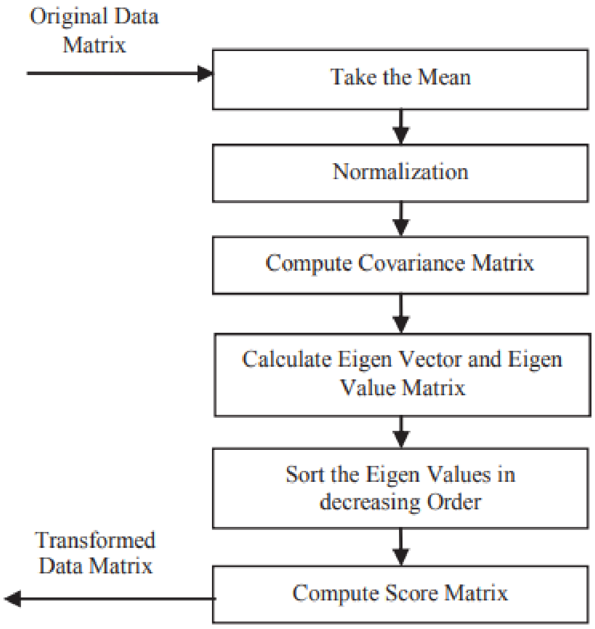
Original Data Matrix
Take the Mean
Normalization
Compute Covariance Matrix
Calculate Eigen Vector and Eigen Value Matrix
Sort the Eigen Values in decreasing Order
Compute Score Matrix
Transformed Data Matrix

Fig. 3. PCA Algorithm flow chart

PCA is an easy-to-use method of dimentionality reduction, which also brings several advantages and disadvantages. For example, because PCA is based on linear algebra, it makes it easy for computers to solve computationally. When machine learning algorithms are trained using principal components rather than the original dataset, they tend to converge more quickly. In context of higher dimensional datasets, regression-based methods are biased towards overfittness of the model. In addition, one can also prevent the prediction or machine learning algorithms from overfitting by applying PCA to reduce the dimensionality of the training dataset beforehand. The significant information loss is one of the key disadvantages of principal component analysis. This is because PCA is designed to find hidden linear relationships between attributes by recognizing orthogonal projections of the data set with the highest variance. For non-linear property of the dataset PCA may produce uninterpreted false positives[24].

4.2 Autoencoders

LeCun proposed the autoencoder concept in 1987. It was the prior method developed to reduce the dimensionality. As the prominence of deep learning research begum, autoencoders are pushed to the forefront of generative modelling. Various academics have suggested numerous autoencoder variations, which have been effectively employed in a variety of domains, including natural language processing, speech recognition, and computer vision [34].

The autoencoder algorithm [33] uses the artificial neural network(ANN) to reduce the dimensionality. The main goal is to reduce the reconstruction error of an input in order to develop a compressed representation of it. For this, Wei Wang1 et al. proposed various types of autoencoders implemented on different dataset to show performance of dimensionality reduction. Recently, the autoencoder method and its variants shows a promising potential to acquire significant characteristics from data, which might lead to the "intrinsic data structure." These techniques, however, only examine self-reconstruction and do not explicitly represent the data relations [35].

4.3 Comparison between Autoencoders and PCA

Even though both the techniques are very useful to resolve curse of dimensionality problem, but has some significant difference based on implementation. Traditional auencoders are based on encoder-decoder architecture to handle dimensionality reduction in unsupervised machine learning to interpret complex non linear functions. Autoencoded features are trained for correct reconstruction, they may have correlations. Additionally, Because of the large number of parameters, autoencoder is prone to overfitting, but such scenarios can be overcome by using regularisation methods. Often, autoencoders are preferred in unsupervised machine learning to identify patterns, image compression for classification, etc. In contrast, PCA is a linear orthogonal transformations which is computationally faster than autoencoders.Because PCA features are projections onto the orthogonal basis, they are completely linearly uncorrelated [36]. It should be also noted here that, for application perspective of PCA, it is very sensitive towards feature value range or scale of feature values. Thus, all the required feature values of an original dataset must be standardised earlier and then PCA could be perform. PCA often proffered in supervised machine learning for dimensionality reduction of numerical features and visualisation of multidimensional data.

5 SUMMARY

As the massive amount of data being generated on daily basis and businesses are becoming more data driven, higher dimensionality problem can not be avoidable in many application domains. There is no such common strategy for massive data anomaly detection. Furthermore, as the volume of data grows, the loss of accuracy becomes larger and the calculation becomes more difficult. High dimensional big datasets can be handled very efficiently by combining various pre-data processing and accessibility tools such as apache pyspark, storm, flink, kafka, MXnets, etc and various ML/DL tecniques. Survey presented here, review the evaluated ways of dealing with the challenge of high dimensionality and offered a overview of few anomaly detection approaches. To overcome the high dimensional data in context of big data, it is necessary that more research and testing of large data anomaly detection algorithms are required.

REFERENCES

[1] Aggarwal, C. C. (2013). Managing and mining sensor data. (Springer eBooks.) New York: Springer.

[2] Jiang F, Leung CK, Pazdor AG. Big data mining of social networks for friend recommendation. In: Advances in social networks analysis and mining (ASONAM), 2016 IEEE/ACM international conference on. IEEE. 2016. pp. 921–2.

[3] Gartner I. Big data definition. https ://www.gartn er.com/it-gloss ary/big-data/. Accessed 14 Feb 2020.

[4] Zhai Y, Ong Y-S, Tsang IW. The emerging "big dimensionality". IEEE Comput Intell Mag. 2014;9(3):14–26.

[5] Zhang L, Lin J, Karim R. Sliding window-based fault detection from high-dimensional data streams. IEEE Trans Syst Man Cybern Syst. 2017;47(2):289–303.

[6] Aggarwal CC. High-dimensional outlier detection: the subspace method. In: Outlier analysis. Springer; 2017. pp. 149–84.

[7] Donoho DL, et al. High-dimensional data analysis: the curses and blessings of dimensionality. AMS Math Chall Lect. 2000;1:32.

[8] Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets. IEEE Trans Knowl Data Eng. 2005;17(2):203–15.

[9] Koufakou A. Scalable and efficient outlier detection in large distributed data sets with mixed-type attributes. Florida: University of Central Florida; 2009.

[10] He Q, Ma Y, Wang Q, Zhuang F, Shi Z, Parallel outlier detection using kd-tree based on mapreduce. In: Cloud computing technology and science (CloudCom), 2011 IEEE third international conference on. IEEE. 2011. pp. 75–80.

[11] Angiulli F, Basta S, Lodi S, Sartori C. Distributed strategies for mining outliers in large data sets. IEEE Trans Knowl Data Eng. 2013;25(7):1520–32.

[12] Bai M, Wang X, Xin J, Wang G. An efficient algorithm for distributed density-based outlier detection on big data. Neurocomputing. 2016;181:19–28.

[13] Sadik S, Gruenwald L. Research issues in outlier detection for data streams. ACM SIGKDD Explor Newsl. 2014;15(1):33–40.

[14] Chu F, Zaniolo C, Fast and light boosting for adaptive mining of data streams. In: Pacific-Asia conference on knowledge discovery and data mining. Springer. 2004. pp. 282–92.

[15] Salehi M, Leckie C, Bezdek JC, Vaithianathan T, Zhang X. Fast memory efficient local outlier detection in data streams. IEEE Trans Knowl Data Eng. 2016;28(12):3246–60.

[16] Gama J. A survey on learning from data streams: current and future trends. Progr Artif Intell. 2012;1(1):45–55.

[17] Yu Q, Tang K-M, Tang S-X, Lv X. Uncertain frequent itemsets mining algorithm on data streams with constraints. In: International conference on intelligent data engineering and automated learning. Springer. 2016. pp. 192–201.

[18] Domingos P, Hulten G. Mining high-speed data streams. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. ACM. 2000. pp. 71–80.

[19] Shin K, Hooi B, Kim J, Faloutsos C. Densealert: Incremental dense-subtensor detection in tensor streams. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM. 2017. pp. 1057–66.

[20] Oh J, Shin K, Papalexakis EE, Faloutsos C, Yu H. S-hot: Scalable high-order tucker decomposition. In: Proceedings of the Tenth ACM international conference on web search and data mining. ACM. 2017. pp. 761–70.

[21] Bellman R. Dynamic programming. Chelmsford: Courier Corporation; 2013.

[22] Thudumu S, Branch P, Jin J, Singh J. Estimation of locally relevant subspace in high-dimensional data. In: Proceedings of the Australasian computer science week multiconference. 2020. pp. 1–6.

[23] Ertöz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings of the 2003 SIAM international conference on data mining. SIAM. 2003. pp. 47–58.

[24] Thudumu, S., Branch, P., Jin, J. et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7, 42 (2020). https://doi.org/10.1186/s40537-020-00320-x.

[25] Koufakou A, Georgiopoulos M. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining Knowl Discov. 2010;20(2):259–89.

[26] Hodge V, Austin J. A survey of outlier detection methodologies. Artif Intell Rev. 2004;22(2):85–126.

[27] Chen G, Iwen M, Chin S, Maggioni M. A fast multiscale framework for data in high-dimensions: measure estimation, anomaly detection, and compressive measurements. In: Visual communications and image processing (VCIP), 2012 IEEE. 2012. pp. 1–6.

[28] Fan J, Fan Y. High dimensional classification using features annealed independence rules. Ann Stat. 2008;36(6):2605.

[29] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst. 1987;2(1–3):37–52.

[30] Shlens J. A tutorial on principal component analysis. arXiv preprint arXiv :1404.1100. 2014.

[31] Chakrabarti K, Mehrotra S. Local dimensionality reduction: a new approach to indexing high dimensional spaces.

[32] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker and Shantanu, "Data analysis using principal component analysis," 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014, pp. 45-48, doi: 10.1109/MedCom.2014.7005973.

[33] D. Rumelhart, G. Hinton, and R. Williams. Learning internal representations by error propagation. Parallel Distributed Processing. Vol 1: Foundations. MIT Press, Cambridge, MA, 1986.

[34] J. Zhai, S. Zhang, J. Chen and Q. He, "Autoencoder and Its Various Variants," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018, pp. 415-419, doi: 10.1109/SMC.2018.00080.

[35] W. Wang, Y. Huang, Y. Wang and L. Wang, "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction," 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 496-503, doi: 10.1109/CVPRW.2014.79.

[36] Muaz, U. (2019, July 25). Autoencoders Vs PCA: When To Use Which ?. Medium. https://towardsdatascience.com/autoencoders-vs-pca-when-to-use-which-73de063f5d7.

[37] Moura, Jose, Serrao, Carlos. (2015). Security and Privacy Issues of Big Data. 10.4018/978-1-4666-8505-5.ch002.

[38] Curse Of Dimensionality Definition | DeepAI. (2019, May 17). DeepAI. https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality.