A Survey on Enhanced Big Sensor Error Detection Model using K – Means Clustering Algorithm

N. Revathi[1] T. P. Senthilkumar[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Gobi Arts & Science College Gobichettipalayam

Abstract— Big Data concern large-volume, complex, growing data set with multiple, autonomous sources. With fast development of networking, data storage and data collection capacity Big Data are now rapidly expanding in all science and engineering domains including physical, biomedical sciences and biological. Wireless Sensor Networks (WSNs) have become a new monitoring solution for a variety of applications and information collections. Error occurring to sensor nodes are common due to the harsh environment where the sensor nodes are deployed and sensor device itself. To ensure the network quality of service it is important for the WSN to be able to detect the errors and take actions to keep away from further degradation of the service. However, these techniques do not give efficient support on quick detection and locating of errors in big sensor data sets. In this paper, develop a new data error detection approach which abuses the full computation capability of cloud platform and the network feature of WSN. That Proposed approach can significantly reduce the time for error detection and location in big data sets created by large scale sensor network systems with acceptable error detecting accuracy. In this paper explores using the K-Mean Clustering algorithm to further improve the accuracy of determining the number of error detection and attackers. In addition, it develops an integrated detection and localization system that can localize the positions of multiple error detection algorithms.

Key words: Big Data, WSN, Sensor Network, Error Detection, Time Efficiency, K-Means Clustering

I. INTRODUCTION

Data mining is a device extracting the data from expansive datasets on the grounds that it is extremely hard to get vital data and give it inside time limit. Data mining [11] is the process of encapsulating into helpful information and analysing data from totally different context. Data mining consists of transforming, extracting, loading transactional data into the data warehouse system, data manage in a multidimensional database system and information storage, it provide data access to presenting the data in a useful format, analysing the data by application software, information technology professionals and business analysts. Data mining includes the clustering, summarization, regression, classification, association rule learning and anomaly detection. In this paper, discuses different clustering algorithm considering the criteria of big data is finished. The expression "clustering" is utilized as a part of a research communities to depict strategies for gathering of unlabelled information [2]. Clustering – is the undertaking of finding structures and groups in the information that are some way or another "comparable", without utilizing referred to structures as a part of the information [8]. The key idea of clustering algorithms is insensitive to the order of input records, ability to deal with noisy data, scalability, etc. Data mining is different kinds of process. It requires evaluating the results and taking relevant action, prospecting the data, converging data for a data mining algorithm and collecting data. The collected data can be stored in operational databases or data warehouse. Data mining is a device extracting the information from large datasets that it is extremely hard to get essential data and give it inside time limit. Big data has typical characteristics of five 'V's, value, validity, velocity, variety and volume [1]. Big data derive by countless areas, including environmental research, gene analysis, biological study, genomics, complex physics simulations, connections, meteorology. Hence, how to process big data due to become a critical challenge for modern society and fundamentals. Cloud computing gives attracted significant attention in alignment with big data, low cost, resource reuse, scalability, storage and platform for big data processing with powerful computation skill. One of the important source for technical big data is too collected by Wireless Sensor Networks (WSN). Wireless sensor networks have idle of expressively upgrading individuals capacity to monitor and cooperate with their physical environment. Big data set from sensors is frequently subject to gift and losses because of wireless medium of explanation and presence of equipment mistakes in the hubs. For a WSN application to understand a suitable result, it is required that the information got is perfect, exact, and lossless. However, cleaning of sensor big data errors is a puzzling issue demanding innovative solutions and effective finding. In data mining clustering that is cluster investigation of information performing main task. Clustering is a method mainly it is difficult at the time of big dataset, group data on basis of their similarities and dissimilarities from data elements. Clustering strategy change over that data into different groups where object in that cluster having properties as compare with other yet not same to different cluster properties. There are different efficient strategies used to take care of the issue for large data clustering. Cluster methods and usage utilized for getting performance and scalability in such information analysis. By utilizing cluster analysis techniques it is anything but difficult to handle complex data sets and K-means is generally utilized for delivering clusters in many application. It is also utilized for compression form and finding some hidden structure and automatically organized data [3] [6] [15]. A WSN is a distributed environment consisting of a substantial number of low-power sensors. In this paper, K-Means clustering algorithm is proposed with the goal of clustering sensors' data in a WSN.

II. ERROR DETECTION

WSN as being hierarchical and heterogeneous. The sensor nodes just disseminate data when the temperature of the

zone being observed goes above or underneath specific limits. Present issue error classification taxonomy for sensors. As of now recommended and utilized one, likewise distinguish a few new types of sensor faults. The best possible beginning stage for testing of sensor-based frameworks is to establish adequate fault models. Error detection models must be sufficiently expressive to appropriately catch every essential features of the most widely common error [5]. In the meantime, they must be computationally and conceptually sufficiently enough so they can be tractable. Distributed sensor-based frameworks are complex frameworks according to at least two criteria. They have a large number of components and interactive system software and they employ a layered and application software component. From the hardware perspective, totally restrict the attention to actuators and sensors. They have a large number of components and they employ a layered and interactive system software and application software components. From the hardware point of view, completely restrict the attention to sensors and actuators. From the functional level perspective, expect that all application software as well as the system software are already fault tolerant. Two types of error classification are extremely mind in terms of their demands how to be treated [6]. The first is one when the sensor readings change is subject to debasement regularly because of the maturing of the sensor. The other requesting model is the one where the sensor readings are changed by a specific function that could possibly be dependent on the sensor readings. For example, on account of light sensors, if there is an error in point estimation, all readings will be changed by a function. The useful technique in fault detection to distinguish a random noise is to run a greatest probability approach on the multi sensor fusion measurements. The random noise would exist, if running these techniques enhances the accuracy of the final results of multi-sensor combination. While there have been a few efforts to minimize irregular mistakes, very little has been done for fault detection. The measurements from multisensory are combined in a model for consistent mapping of the sensed phenomena. Although the fact that the new fault detection technique is nonexclusive and can be connected to an arbitrary system of sensors that utilization an arbitrary type of data fusion. In a various hierarchical system, hubs are grouped into clusters and there is a special hub called cluster-head. In a heterogeneous system, the cluster-heads have more assets and therefore are more effective than the common-hubs. Besides, they are responsible for sending information to a base station (BS). The BS additionally communicates with the observer, which is a system substance or a final client that needs to have data about information gathered from the sensor hubs. In this paper discuss about the clustering aims to naturally group related error into single clusters and also presents a new spectral clustering method called Correlation Preserving Indexing (CPI), which is performed in the relationship similarity measure space. In this research, the error detection are view into a low-dimensional semantic space in which the connections between the error detection in the local fixes are maximized while the relationships between the error identification outside these patches are minimized at the sometimes.

III. RELATED WORK

Mittal Namita proposed in their paper ―Hybrid Approach for Detection of Anomaly Network Traffic utilizing Data Mining Techniques a cross breed approach that endeavours the advantages of both the strategies i.e. entropy based and bolster vector machine based individually. Mixture abnormality location framework takes in the conduct of system activity from the standardized entropy estimations of various system features [7]. Entropy based procedures have the upside of better speaking to the properties of the system movement and reinforce vector machine is useful for arrangement. The standardized entropies are sent to SVM model for taking in the conduct of the system. This prepared SVM model can arrange the system movement in assault activity or real movement. In entropy based irregularity identification framework, firstly standardized entropy of system activity components is figured in like clockwork. Limit quality is altered for every component for distinguishing the oddities in view of analyses. At that point voting framework for every component chooses whether there is an assault or not. This strategy can deliver great results if there should be an occurrence of recognizing assault movement however it additionally creates high false alerts, in light of the fact that the entropy qualities can likewise veer off from the reach or towards 0 or 1 in the event of authorized traffic.

Mr Jain proposed in their paper a cross breed display that coordinates Anomaly based Intrusion recognition strategy with Signature based Intrusion discovery procedure which is separated into two phases [1]. In first stage, the mark based IDS SNORT is utilized to produce cautions for abnormality information. In second stage, information mining procedures "k-implies + CART" is utilized to course k-implies grouping and CART (Classification and Regression Trees) for characterizing ordinary and unusual exercises. The half and half IDS model is assessed utilizing KDD Cup Dataset. The proposed array is acquainted with augment the viability in recognizing assaults and accomplish high exactness rate and in addition low false caution rate [1].

Krishnamachari and Iyengar propose a few limited edge based choice plans to recognize both defective sensors and occasion locales. The 0/1 choice predicates from the area are gathered and the quantity of neighbours with the same predicates are computed. This number is utilized for a definite choice taking into account a dominant part vote [4]. The calculation introduced in this paper recognizes both defective sensors and cutting edge of occasion districts. It functions admirably with 0/1 decision predicates as well as numbers that dynamic sensor readings or sensor practices.

Ramanathan Clouqueur, Saluja, and seek algorithms to collaboratively detect the presence of a target in a region G. Each sensor obtains compares it with a pre-determined threshold for final decision, target energy (or local decision) from all other sensors in the region and then drops extreme values if faulty sensors exist. Correspondingly, the algorithm is termed value combination if the information is target energy or decision combination if the info is the local decision made by each sensor [13]. Under these algorithms, all sensors in region G will close with a same choice. Be that as it may, both target energy and local choice should be processed early. Further, these

algorithms don't indicate how to characterize locale G, as the correspondence and algorithm overheads are strongly identified with the size of G. The algorithm requires just raw readings from the area and catches the limit sensors surrounding the region.

Nadianmai G. V and M. Hemalatha in their paper using data mining techniques considered four issues namely Effectiveness of Distributed Denial of Service attack, Lack of Labelled Data, High Level of Human Effective approach toward Intrusion Detection System Interaction, Classification of Data and solved them using the proposed algorithms Semi-Supervised Approach and Varying HOPE RAA algorithm, EDADT algorithm, Hybrid IDS model algorithm respectively. To solve the problem related to an enhanced data adapted decision, classification of data [9].

V. Barot and D. Toshniwal discussed a hybrid model that troupes Naive Bayes (factual) and Decision Table Majority (principle based) approaches [14]. Naive Bayes predicts rapidly in light of less intricate working of it and procedures preparing information set just once to store measurements. Decision Table Majority (DTM) is a classifier that matches each of the property values all together. This model uses successive renaming approach for joining guideline base classifier. Here CFS algorithms is utilized for property choice utilizing Best First search. Author utilized KDDCUP'99 information set for their experiments.

Wankhade K, Patka S, Thool R discussed hybrid data mining approach enveloping clustering ensemble, merge and divide, clustering feature selection, filtering [5]. A methodology for assessing the number of the centroid and selecting the appropriate early cluster centroid is displayed.

Dhakar M, Tiwari A, in context to upgrade execution, the work shows a model for IDS. This enhanced model, named as REP (Reduced Error Pruning) based IDS Model gives yield with greater perfection alongside the expanded number of appropriately ordered occurrences [6]. It utilizes the two algorithms of classification methodologies to be specific, K2 (BayesNet) and REP (Decision Tree). Here REP gives a successful characterization alongside the pruning of tree with choice learning ability.

Subramanian P.R and Robinson J.W discussed on network security through Intrusion Detection Systems (IDSs) with information mining approaches [12]. This model uses twofold multi boosting technique strategy and binary classifier (C4.5) [9]. Here binary classifier is utilized to reduce the variance and bias multi boosting technique is used and used for each type of attack to improve the accuracy, to classify bit by bit transmission of the packet.

N.S and Nandavadekar V.D introduce a methodology for intrusion detection using J48 decision tree classifier furthermore contrasted and some other tree based algorithms in which J48 tree demonstrates the best execution [10]. To assess the execution of the algorithm effectively Kappa statistics measures, Root Relative Squared Error, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), classified instances.

IV. Conclusion

Big Data as a developing pattern and the requirement for Big Data mining is emerging in all engineering domains and science. Big Data technologies gives a relevant and most accurate social sensing information to better understand society at real-time. The time of Big Data has arrived. In large volume of digital information is created by many applications. Clustering is broadly utilized technique intended for data mining tool and knowledge discovery. Clustering assumes extremely crucial part in data mining and utilized by tool for big data analysis. In this research elaborate the problem of data clustering in sensor network. The proposed K—Means clustering algorithm to deal with clustered network of sensor and different points of utilizing sensor platform as a part of error detection process. Incorporating sensor platform for error detection process in complex system frameworks made secure, efficient and lossless transmission in timely manner.

Reference

[1] A.K. Jain, M.N. Murty & P.J. Flynn, “Data Clustering: A Review”, ACM Computer Survey, vol. 31, no. 3, pp. 264–323, 1999.
[2] Avita Katal, Mohammad Wazid, and RH Goudar, “Big data: Issues, challenges, tools and good practices”. In Contemporary Computing (IC3), Sixth International Conference on, pages 404-409. IEEE, 2013.
[3] Fahim, A. M., A. M. Salem, F. A. Torkey, and M. A. Ramadan. "An efficient enhanced k-means clustering algorithm." Journal of Zhejiang University SCIENCE A 7, no. 10, 1626-1633, 2006.
[4] B. Krishnamachari and S. Iyengar, Distributed Bayesion Algorithms for Fault- Tolerant Event Region Detection in Wireless Sensor Networks, IEEE Transactions on Computers, Vol. 53, No. 3, pp. 241 – 250, March 2004.
[5] K. Wankhade, S. Patka and R. Thool, “An efficient approach for Intrusion Detection using data mining methods”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), Print ISBN:978-1-4799-2432-5 INSPEC Accession no. 13861274, August 22-25, pp. 1615-1618, 2013.
[6] M. Dhakar and A. Tiwari, “A New Model for Intrusion Detection based on Reduced Error Pruning Technique” International Journal of Computer Network and Information Security, pp. 51-57,2013.
[7] Mittal Namita, - Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Technique, Elsevier, Procedia Technology 6,996-1003,2013.
[8] M.Vijayalakshmi, M.Renuka Devi, “A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets” ,International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.
[9] Nadiammai G. V., Hemalathain M.,—Effective approach toward Intrusion Detection System using data mining techniques, Cairo University, Elsevier, Egyptian Informatics Journal, pp. 37-50,2014.
[10] N. S. Chandolikar and V. D. Nandavadekar, “Efficient algorithm for intrusion attack classification by analysing KDD Cup 99”, Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on ISSN :2151-7681, September 20-22, pp. 1 – 5, 2012.

[11] Oded Maimon, Lior Rokach, "Data Mining AND Knowlwdge Discovery Handbook", Springer Science+BusinessMedia.Inc, pp.321-352, 2005.
[12] P. R. Subramanian and J. W. Robinson, "Alert over The attacks of data packet and detect the intruders", Computing, Electronics and Electrical Technologies (ICCEET), IEEE International Conference on ISBN: 978-1-4673-0211-1, pp. 1028-1031,2012.
[13] T. Clouqueur, K.K. Saluja and P. Ramanathan, Fault Tolerance in Collaborative Sensor Networks For Target Detection, IEEE, Transactions on Computers, pp. 320-333, Vol. 53, No. 3, March 2004.
[14] V. Barot and D. Toshniwal, "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE International Conference on Print ISBN: 978-1-4673-2148-8, July 18-20, 2012.
[15] Yugal Kumar, Yugal Kumar, and G. Sahoo G. Sahoo. "A New Initialization Method to Originate Initial Cluster Centres for K-Means Algorithm." International Journal of Advanced Science and Technology 62, 43-54, 2014.

