A Review on Efficient Approaches to Detect and Eliminate Data Redundancy in Large Volume of Data using Anomaly detection

[1]Mr.T.Somashekar, [2]Dr.A.Arun Kumar
[1]Research Scholar, [2]Professor
Computer Science & Engineering
Balaji Institute of Technology &Science Warangal, India

Abstract : In order to finding an un-matching pattern in any dataset that does not satisfy the expected nature of the customer then Anomaly detection will these kinds of issues and Anomaly detection also finds the inconsistent data pattern, and this process is called as novelty detection, noise mining, and anomaly mining. Modern IT companies enable enterprises to detect strange events automatically in streaming data. Un-matching pattern refers error in the dataset, different pattern, duplicate data and misbehavior data. Identifying anomalies is more important in a wide range of disciplines like economic data, medical analysis, share market, insurance data and identity fraud, network malicious and programming defects. There are various types of anomalies available such as point or content anomalies, context anomalies, and collective anomalies. Some of the data are abnormal than the other entire dataset regarding meta-information is called as context anomalies. The collected data points are considered as anomalies when compared to other data in the data sheet.

Keywords: Data Mining, data preprocessing, Big data, MOMGODB, QAmodel

I. INTRODUCTION:

In general, anomaly detection can be obtained by three types of algorithms such as unsupervised, supervised and semi-supervised algorithms. These algorithms utilize labeling the trained data and compare with the test data. To separate normal and abnormal data, labeling and comparing are used. This classification of training data leads to analyze the new entry test data while streaming. Various issues to be challenged with standard anomaly detection methods due to the fields like spatial, sequential or temporal data associated with the sources from where the data are generated. In recent days anomaly detection is used mainly for prevalent Big data especially sensor data. Sensor data are recorded remotely using various sources such as electrical outlets, weblogs, water pipes, telecommunications and many other areas.

Compared with a template of large amounts of data which is input very frequently. Anomaly detection is also a kind of intrusion detection method. Digital media and its contents are increasing tremendously that creates a challenging problem for data administrators. Shaping and organizing data from various resources to the data repositories are based on the schema and structure of the data. This arrangement can be made by some set of software agents installed in the digital libraries. If the size of the data increases then it is hard to manage the entire dataset and problems occur regarding response time, availability, security and quality assurance. To improve the peculiarity mining, the dirty data (i.e. replicas, errors and unique patterns) from the repositories should be removed.

Data Mining

Data mining has become one of the most promising and progressive fields for the manipulation and extraction of data to produce useful information. Every day most of the businesses are using data mining applications to extract, manipulate, and identify valuable information from the records stored in their databases, data warehouses, and data repositories. Process optimization, human factors, shop scheduling, and quality management are some of the areas in which data mining tools are used such as decision trees, genetic algorithms, data visualization, and neural networks can be implemented with great results.

Anomaly

Identifying abnormal data points, data items, events which cannot imitate to the predictable design of a given data group. These are some of the anomalies which are created not frequently, but it is a significant threat like fraud or cyber intrusions. Anomaly detection is applied in behavioral analysis and another format of examination to aid in learning about location, detection, identification and predicting of anomalies in the large set of data. It is also called as anomaly detection or outlier detection.

1.1 Anomaly Detection

One of the most important processes used as the main process in data mining is anomaly detection. It is used to determine the kind of anomalies exists in a given dataset with the details of their creation. Fraud detection, fault detection, intrusion detection, event detection, health care monitoring kinds of domains needs anomaly detection. Mostly in sensor networks, nomaly detection is used widely. Anomalies are unexpected abnormal activities occur whereas it is detected by fetching the abnormalities of the data, data behavior, and other rare abnormal activities. Anomaly detection is most important since it destroys the quality of data and data mining process. Anomaly detection is a problem of finding errors, different patterns, duplication, and misbehavior. One of the major research problem based on applications is anomaly detection. Various non-conforming patterns, aberrations, peculiarities or exceptions in various application domains are often referred to as anomalies. From these, the duplicate records, error, misbehavior based data are treated as anomalies. Anomaly detection is mostly used in online applications such as banking, credit card fraud, insurance, and healthcare.

Anomaly detection is also a kind of intrusion detection method. Anomaly detection is as follows

- Identifying the unknown data unknowingly.
- Providing alert message about anomalies to the appropriate users, where it leads to classifying the entire document according to the relevancy.
- It is also sharing knowledge about anomalies, which helps the organization as a whole can better understand how to handle it.

Classification of Data Anomalies

- Data anomalies are characterized by semantic, syntactical and coverage anomalies.

Syntactical Anomalies

- Lexical errors

Semantic Anomalies

- Integrity constraint violations
- Contradictions
- Duplicates
- Invalid tuples

Coverage Anomalies

- Missing values
- Missing tuples

Various Kinds of Anomaly Detection Techniques

Much earlier research works proposed for data cleaning and data preprocessing methods. Mainly supervised, unsupervised and machine learning approaches are used to do data preprocessing jobs in database oriented applications. RajuDara et al. (2015) proposed a framework which implements sound data quality to ensure consistent and correct loading of data into data warehouses which ensure accurate and reliable data analysis, data mining and knowledge discovery.

Sapna Devi et al. (2015) discussed an overview of data cleaning problems, data quality, cleaning approaches and comparison of data cleaning tool. This tool does auditing on the data to find the types of anomalies contained within the data. From the above literature survey, it is clear that it is very much important to preprocess the data to detect and eliminate the errors and redundancy of the data.

Gabriel Poesia et al. (2014) describe the patterns in the dataset; N-ary relations have been computed where it provides relations in one-dimensional data. DBLEARN and DBDISCOVER are two systems developed to analyze RDBMS. Weyuker, E. J. et al. (2000) the worth of the software products depends on the data models used and dynamic changes in the data.

Chu et al. (2013) the interaction of the different constraints by encoding them in a conflict hyper graph and have shown that this approach to holistic repair improves the quality of the cleaned database on the same database treated with a combination of existing techniques.

Jeby K Luthiya et al. (2013) PSO algorithm is used to control the redundancy and duplication. The proposed PSO generates optimal similarity measures to decide whether the data duplicated in the training dataset or not.

Donghun Lee et al. (2012) described a framework that has been developed to manage performance anomalies after establishing a set of conditions for a problem to be considered an anomaly. The framework uses Statistical Process Control (SPC) charts to detect performance anomalies and differential profiling to identify their root causes. By automating the tasks within the framework, which able to remove most of the manual overhead in detecting anomalies and reduce the analysis time for identifying the root causes by about 90 percent in most cases.

Deepa, K. et al. (2012) proposed a heuristic global optimization method called Particle Swarm Optimization algorithm for record de-duplication. They considered the fitness function of the PSO algorithm, and it based on the swarm of data. Here the proposed approach has two phases such as training phase and the duplicate detection phase. First, they find the similarity between all attributes of record pairs using Levenshtein distance and cosine similarity. Then they formed the feature vectors for representing the set of elements which required detection of duplicates. From this feature vectors, they found the duplicate records by using the PSO algorithm.

Vikrant Sabnis et al. (2012) the chief objective of the data mining technique is to sense and classify data in a huge set of a database without negotiating the speed of the process. PCA used for data reduction, and SVM used for data classification in.

Kumbhar and Krishnan (2011) have been presented an Artificial Bee Colony (ABC) based methodology, which maximizes its accuracy and minimizes the number of connections of an Artificial Neural Network (ANN) by evolving at the same time the synaptic weights, the ANN's architecture and the transfer functions of each neuron. The methodology tested with several pattern recognition algorithms.

Ireneusz Czarnowski et al. (2008) described an approach to data reduction. For machine learning and data mining, this data reduction functions are very vital. For solving data reduction, an agent based population algorithm used. Data reduction is not only the solution for improving the quality of databases. Different sizes of the database are used to provide high classification among the data to find out anomalies.

Haidarian et al. (2006) when the size of the data increased then the number of computational resources also gets better; also disconnecting duplicate record is also tough due to the scale of the database.

Kordas, N. et al. (2006) discussed the existing approaches presented for data cleaning to develop DBMS applications for deciding the problem statement, different anomaly detection methods have been proposed where some of them are application specific, and others are generic. The process of sensing and removing the duplicate record in a repository is called record de-duplication.

Jose Ramon Cano et al. (2005) two algorithms such as evolutionary and non-evolutionary are applied, and the results are compared to find the best suitable algorithm for anomaly detection.

Thomas Zimmermann et al. (2004) discussed the data preprocessing and data manipulations in four stages. A grained analysis of CVS archives is applied regarding data extraction, transaction recovery, mapping of changes to fine-grained entities and cleaning the data. The data cleaning methodology is used as a building block and inserted into the functional blocks of the program.

Denaro, G. et al. (2004) the DBMS engine or each release is required to have a grand deployment of up-gradation of the application software. Moreover, the data used in the application software depends on the hardware technology also. Searching the difficulties and the reasons for the problems should be acknowledged in all application oftware with the help of testing techniques which is already available. Also, the testing process applied at the end of the deployment process.

Wheatley, M. et al. (2004) two different datasets found in this approach are better and provide 6.2% of accuracy more than the earlier approaches. It can extend for different benchmark data with real-time data such as time series data, clinical data, 20-20 new group, etc.

Dutch et al. (2011), Dubnicki et al. (2009), Ungureanu et al. (2010), Bolosky et al. (2000) in the file system, there is an enormous set of file information, and more copies of information stored. Same files, sub-files, sub-file regions persisted in the same system or different system in a network. For improving the high space formation in a file system, this duplicating storage should be de-duplicated. The sub-directory or in the whole directory is allotted for de-duplication to work.

All the data used in the industry categorized according to the basic data units used to handle the data. In the present approaches, there are two main methods that is used for de-duplication; one is File-level de-duplication Harnik, D. et al. (2010), Gunawi, H. S. et al. (2005), Douceur et al. (2002) and the other one is block-level de-duplication Quinlan et al. (2002), Muthitacharoen et al. (2001), Vrable et al. (2009). File-level de-duplication examines the external information of the file system whereas the block level de-duplication investigates the internal information of the file system or the data in the storage. Due to the new incoming data, there are three major issues are to be faced; they cannot buy more storage; take the longer time to make the back-up of the data and takes the longer time to recover the data also. Each group needs to run their application software without any interruption during the operation. So, it is necessary to deploy the data with better quality, and it can provide the best service to the users.

Rita Aceves-Pérez et al. (2005) software developed with safe, correct and reliable operations for avionics and automobile based database systems. To keep away from web-based data redundancy. A statistical QA model is applied to develop a prototype GDW, ThiagoLuís Lopes Siqueira et al. (2009), SOLAP is functional to Gist database and other spatial database analysis, indexing, and generating various set of reports without any error.

Ahmad Ali Iqbal et al. (2010) an effectual method has proposed for point to point sharing data. During the data sharing, the data duplication is removed using the efficient method. YanxuZhu et al. (2011) web entity data extraction connected with the attributes of the data can be received using a novel approach which uses duplicated attribute value pairs.

Ye Qingwei et al. (2010) eliminating the record duplication is the chief aim of the current approach. The experimental results obtained from the modern approaches PSO and GA are compared to evaluate the performance, where PSO is better than GA is proved.

Weyuker, E. J. et al. (2000) anomaly detection methods used in various applications like online banking, credit card fraud, insurance and health care areas. The quality of the software product is depending on the data models used for running dynamic changes on the data. Some of the research works do testing development also integrating with the anomaly detection Denaro, G. et al. (2004).

Few research works are removing the duplicate record in file system either at sub-directory or in the whole directory Dutch et al. (2011), Dubnicki et al. (2009), Ungureanu, C. et al. (2010), Bolosky, W. J. et al. (2000). These existing approaches mainly divided into two types of categories like de-duplication, one is file-level Harnik, D. et al. (2010), Gunawi, H. S. et al. (2005), Douceur et al. (2002) and the other one is Block-level de-duplication Quinlan et al. (2002), Muthitacharoen et al. (2001), Vrable et al. (2009). It means that the duplication records analyzed regarding internal information and external information of the file system. From this point of views, it has been decided that it is essential to develop a method for deleting data duplication to increase the quality of the data from the above background study.

1.2 Survey on Data Duplication Removal

Peter Christen (2012) surveyed various indexing techniques for record linkage and de-duplication. Record linkage refers to the task of identifying records in a dataset that relates to the same entity across different data sources. The blocking technique used in traditional record linkage approach. Blocking key values are used to place the files into different blocks.

Ahmed, K. et al. (2007) many problems expected while integrating data from various sources into a single database. These troubles occur due to the heterogeneity of the data.

Moises, G. et al. (2006) proposed a method to confirm de-duplication should complete individual but contradictory objectives: this process should effectively increase the identification of the records replicated. For searching accurate answers to a given problem without searching all the data, on the whole, it is apt to have the basic approach, known as the GP. Because of the record de-duplication problem, in the current approach.

Moises, G. et al. (2008) without PSO based approach, the present system results took for comparison, where the approach can find more efficient de-duplication methods without human intervention. Also, PSO based approach can be interoperable with existing best de-duplication methods to change the replication identification limits used to categorize a pair of records as a match or not. From the data content, several different portions of attestation are extracted for the requirement to exhibit a de-duplication function, that is capable of recognizing whether two or more entries in a database are replicas or not. This approach binds several different portions of attestation extracted from the database. It is to record that de-duplication is a time-consuming work even for small databases. The scope is to support a process that predicts a peculiar combination of the best pieces of evidence and so when yielding a de-duplication function. Further, it improves the performance using a method to compare the corresponding data for the training process.

Even applied to other databases with similar characteristics, this function can apply on the entire data at last. There is no unexpected deviation in the data patterns, and recent supplemental data can be entreated similarly by the suggested function, as long as something that is very significant in huge databases. It is valuable concern that these (arithmetical) services, which can consider as a combination of numerous powerful de-

duplication regulations, is easy, fast and vigorous to calculate, allow its capable technique to the de-duplication of huge databases. For registering de-duplication Moises, G. et al. (2012), a Hereditary Scheme (HS) approach used. The problem of finding out and destroy replica entries in a repository is known as record de-duplication Kordas, N. et al. (2006).

Using HS approach, the genetic operation is carried out by record de-duplication that generates gene excellence for each record. If that gene value is the same as any other record, that record considered as a duplicate record. These trading operations are to increase the characteristic of given record. Genetic Operations are Reproduction, Mutation, and Crossover. From this, it can understand that genetic operations could blow the performance of the record de-duplication task. From the experimental results, it is inferred that the significant difference between the different efforts is required to obtain a quick solution.

Moises, G. et al. (2012) suggested an approach for record de-duplication by applying the genetic programming. Using genetic operations, genetic programming approach is introduced to remove the de-duplication where it produces gene values which signify data records. Record duplication is nothing but the matching of gene value from one record with another record. The GA approach to record de-duplication is to combine l different pieces of evidence extracted from the data content and to devise a de-duplication function that will be able to identify whether two entries in a data store are replicas or not. This operation can be used to enhance the attributes of each record in a database.

CONCLUSION:

The main objectives of this research work are to provide better data preprocessing methodology regarding error removal and redundancy reduction to increase the efficiency of the data quality for data mining. In this research, there are three different stages are applied using three different solutions.

REFERENCES

1. Aceves-Pérez, Rita., Villaseñor-Pineda, Luis and Montes-y-Gomez, Manuel "Towards a Multilingual QA System Based on the Web Data Redundancy", International Atlantic Web Intelligence Conference, Springer, pp.32-37, 2005.

2. Agarwal, D. "An Empirical Bayes Approach to Detect Anomalies in Dynamic Multidimensional Arrays", In 5th IEEE International Conference on Data Mining, IEEE Computer Society, pp.26-33, 2005.

3. Agyemang, M., Barker, K. and Alhajj, R. "A Comprehensive Survey of Numeric and Symbolic Outlier Mining Techniques", Intelligent Data Analysis, Vol.10, pp.521-538, 2006.

4. Ahmed, K., Ipeirotis, G. Panagiotis, and Verykios, S. Vassilios "Duplicate Record Detection: A survey", IEEE Transactions On Knowledge And Data Engineering, Vol.19, No.1, 2007.

5. Bolosky, W.J., Corbin, S., Goebel, D. and Douceur, J.R. "Single Instance Storage in Windows® 2000", In 4th Conference on USENIX Windows Systems Symposium, USENIX Association Berkeley, CA, USA, pp.2-2. 2000.

6. Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleyner, G., Mueller, M., Singhal, S. and Cohen, I. "Self-aware Services: using Bayesian Networks for Detecting Anomalies in Internet-based Services", Integrated Network Management, IEEE/IFIP International Symposium, 2001.

7. Cano, Jose Ramon., Herrera, Francisco and Lozano, Manuel "Strategies for Scaling Up Evolutionary Instance Reduction Algorithms for Data Mining", Evolutionary Computation in Data Mining, Springer-Verlag, Vol.163, pp.21-39, 2005.

8. Ahamed, B. B., & Hariharan, S. (2012, December). State of the art process in query processing ranking system. In 2012 Fourth International Conference on Advanced Computing (ICoAC) (pp. 1-5). IEEE.

9. Christen, Peter. "A Survey of Indexing Techniques for Scalable Record Linkage and De-duplication", IEEE Transactions on Knowledge and Data Engineering, Vol.24, No.9, pp.1537-1555, September 2012.

10. Collet, Christine., Huhns, N. Michael and Shen, W.M. "Resource Integration using a Broad Knowledge base in Carnot", IEEE Computer Society Press Los Alamitos, CA, USA, Vol.24, pp.55-62, 2002.

11. Ahamed, B. B., & Hariharan, S. (2012). Integration of Sound Signature Authentication System. International Journal of Security and Its Applications, 6(4), 77-86.

12. Damasio, C.V., Analyti, A., Antoniou, G. and Wagner, G. "Supporting Open and Closed World Reasoning on the Web", Principles and Practice of Semantic Web Reasoning (PPSWR06), Springer, pp.21-36, 2005.

AUTHORS BIBLIOGRAPHY

T.Somashekar, Research Scholar at Osmania University and worked as Asst.Professor in BITS engineering college Warangal and area of interest in image processing, data science.

Dr.A.Arun Kumar, presently working as Professor in BITS engineering college Warangal and area of interest in image processing, Network Security and data science