Novel Approach for Outlier Detection Technique using Property Pair Extraction Algorithm

C. Prakash

Assistant Professor, Department of Information Technology, Dr. N.G.P. Arts & Science College, Coimbatore.

A B S T R A C T

Outliers are extreme values that deviate from other observations on data; they may indicate variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. Challenges in anomaly detection include appropriate feature extraction, defining normal behaviors, handling imbalanced distribution of normal and abnormal data, addressing the variations in abnormal behavior, sparse occurrence of abnormal events, environmental variations, camera movements, etc. With a view to successful outlier detection, the proposed PEP algorithm applies a provisional model that recognises an exceptional property par with the most suitable method of implementation. Several outlier identification methods have been developed for some domains and applications, but the approaches have been more general and are subject to confidentiality problems. The proposed concept essentially applies the Genetic Modal Based Approach, which is called the GENEX algorithm and the PEP algorithm for the identification of sub-population scores for both numerical and categorical datasets. In addition, the system performs the best fit method to find the best class based on score and mark. The proposed algorithm will minimise the cost of computing and the lack of accuracy of the problem by applying the best data mining and appropriate pruning techniques. Experiments and outcomes have the best match values for the moderate and extreme outlier ranges.

Keywords: Outlier, Detection, Generic Approach Image Processing, Imaging Systems, Pattern Recognition, Optical Security and Encryption, K-means algorithm, FP-Growth algorithm, Direction Lower Triangular.

1. Introduction

Anomaly detection (also known as outlier detection) is the search for data items in a dataset which do not conform to an expected pattern. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. Anomalies are also referred to as outliers. Outlier detection is the process of identifying abnormal pattern from set of objects. Outlier detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of "outlier" is given in [1] "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism," which gives the general idea of an outlier and motivates many anomaly detection methods. Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behavior. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains [2].

Outlier detection can usually be considered as a pre-processing step for locating, in a data set, those objects that do not conform to well-defined notions of expected behavior. It is very important in data mining for discovering novel or rare events, anomalies, vicious actions, exceptional phenomena, etc.

* *Corresponding author*.
E-mail address: profksiva@gmail.com

2. Related Work

There are several approaches are identified in this literature review for solving this research problem. But each approach has its own features and shortcomings also. This section summarises the literature review of the issue.

Karanjit Singh, Shuchita Upadhyaya (2012) proposed, Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques ha ve been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. The techniques and results of such techniques are not readily forthcoming. A number of surveys, research and review articles and books cover outlier detection techniques in machine learning and statistical domains individually in great details. In this paper we make an attempt to bring together various outlier detection techniques, in a structured and generic description [14].

Attila Tiba; Zsombor Bartik; Henrietta Toman; Andras Hajdu (2019) proposed, there are numerous reasons why inappropriate data can occur in a database. It is essential to detect and eliminate these elements for getting accurate results and conclusions. The process of filtering anomalies generated in the database is called outlier detection. The outliers, that are extreme values deviating from other observations on data, indicate variability in a measurement, experimental errors or a novelty. In this paper, an ensemble-based outlier detection method is presented, where the members of the ensemble are convolutional neural networks (CNNs) combined with a Support Vector Machine (SVM) classifier. Ensemble-based methods are highly popular approaches that increase the accuracy of a decision by aggregating the opinions of individual voters. From hybrid models constructed by the neural networks and the SVM classifier, we have created an ensemble system that makes decision based on the conception of majority voting and results in very accurate outlier filtering. It actually serves as a pre-filter that can be integrated into a more exhaustive image analysis process with rejecting images that fall outside the domain or have poor quality [15].

Cai Lile; Li Yiqun. (2017) proposed, Infrared thermography has become an effective tool in electrical preventive maintenance program due to its high precision and the capability of performing non-contact diagnostic. Anomalies in a thermal image are typically detected by comparing the temperatures of the equipment with reference temperatures. Manual detection is time-consuming and unreliable, making it unable to meet the excessive demand for condition monitoring in industrial applications. In this paper, we propose an automatic method to detect thermal anomalies based on deep neural networks (DNNs). The DNN model is trained to learn the statistical regularities of normal thermal images, and anomalies are detected based on pixel-wise comparison between the learned reference temperatures and the actual temperatures. We test our method on a variety of electri cal equipment and the experimental results demonstrated the effectiveness of the proposed method [16].

Charlie Isaksson, Margaret H. Dunham (2009) proposed, data Mining is the process of extracting interesting information from large sets of data. Outliers are defined as events that occur very infrequently. Detecting outliers before they escalate with potentially catastrophic consequences is very important for various real life applications such as in the field of fraud detection, network robustness analysis, and intrusion detection. This paper presents a comprehensive analysis of three outlier detection methods Extensible Markov Model (EMM), Local Outlier Factor (LOF) and LCS-Mine, where algorithm analysis shows the time complexity analysis and outlier detection accuracy. The experiments conducted with Ozone level Detection, IR video trajectories, and 1999 and 2000 DARPA DDoS datasets demonstrate that EMM outperforms both LOF and LSC-Mine in both time and outlier detection accuracy [17].

Viswanath, S., & Madabhushi, A. (2012) proposed, dimensionality reduction (DR) enables the construction of a lower dimensional space (embedding) from a higher dimensional feature space while preserving object-class discriminability. However several popular DR approaches suffer from sensitivity to choice of parameters and/or presence of noise in the data. In this paper, we present a novel DR technique known as consensus embedding that aims to overcome these problems by generating and combining multiple low-dimensional embeddings, hence exploiting the variance among them in a manner similar to ensemble classifier approaches such as Bagging. We demonstrate theoretical properties of consensus embedding which show that it will result in a single stable embedding solution that preserves information more accurately as compared to any individual embedding and code parallelization are utilized to provide for an efficient implementation of the approach [18].

Xiaodan Xu, Huawen Liu, and Minghai Yao (2018) proposed, anomaly analysis is of great interest to diverse fields, including data mining and machine learning, and plays a critical role in a wide range of applications, such as medical health, credit card fraud, and intrusion detection. Recently, a significant number of anomaly detection methods with a variety of types have been witnessed. This paper intends to provide a comprehensive overview of the existing work on anomaly detection, especially for the data with high dimensionalities and mixed types, where identifying anomalous patterns or behaviours is a nontrivial work. Specifically, we first present recent advances in anomaly detection, discussing the pros and cons of the detection methods. Then we conduct extensive experiments on public datasets to evaluate several typical and popular anomaly detection methods. The purpose of this paper is to offer a better understanding of the state-of-the-art techniques of anomaly detection for practitioners [18].

3. Application of Outlier Detection

Outlier identification has become a well-researched issue and is commonly used in a wide range of application areas such as credit card, insurance, tax fraud detection, cyber security intrusion detection, security vital device fault detection, military monitoring of enemy activities and many other areas[3]. Finding outlier subpopulation is a time-consuming process; the existing system only defines irregular behaviour as outlier rather than range estimation.

3.1. Existing Approach

Several clustering techniques have been applied Clustering algorithms, which are optimized to find clusters rather than outliers. So that produced the following basic problems.

- Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters.

A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers [4]. The existing system discovers attributes or properties based on the given populations which are called as inliers. Existing methods are,

- **SVM based outlier detection**
- **Probabilistic model**
- **Statistical modal**
- **Distance based approaches**

EXPREX algorithm: Finally a sub population creation method for identifying exceptional property, EXPREX algorithm has been applied. Drawbacks are,

- Need more inliers and not suitable for high dimensional datasets.
- Uses only available datasets rather than predicting next points.

Ineffective when high dimensional dataset given. Cost and time delay.

4. Proposed Methodology

The proposed system implements genetic approach and a PEP algorithm and extends the perspective of that approach in order to be able to deal with groups, or subpopulations, of anomalous individuals. As an example, consider a rare disease and assume a population of healthy and unhealthy human individuals is given; here, it would be very useful to single out properties characterizing the unhealthy individuals [5]. An exceptional property is an attribute characterizing the abnormality of the given anomalous group (the outliers) with respect to the normal data population (the inliers). If the inliers data"s are not much sufficient, then the system will analyze and cross check the available dataset for further process.

Moreover, each property can have associated a condition, also called explanation, whose aim is to single out a (significant) portion of the data for which the property is indeed characterizing anomalous subpopulations. In order to single out significant properties, this resorts to minimum distance estimation methods that are statistical methods for fitting a mathematical model to data.

Additionally the system implements the LOO strategy the system will finds the principal direction for the outlier detection. The proposed system observes that removing (or adding) an abnormal data instance will affect the principal direction of the resulting data than removing (or adding) a normal one does. Using the above "leave one out" (LOO) strategy, they can calculate the principal direction of the data set without the target instance present and that of the original data set [6].

By ranking the difference scores of all data points, one can identify the outlier data by a predefined threshold or a predetermined portion of the data. They note that the above framework can be considered as a detrimental principal component based approach for outlier detection.

While it works well for applications with moderate data set size, the variation of principal directions might not be significant when the size of the data set is large.

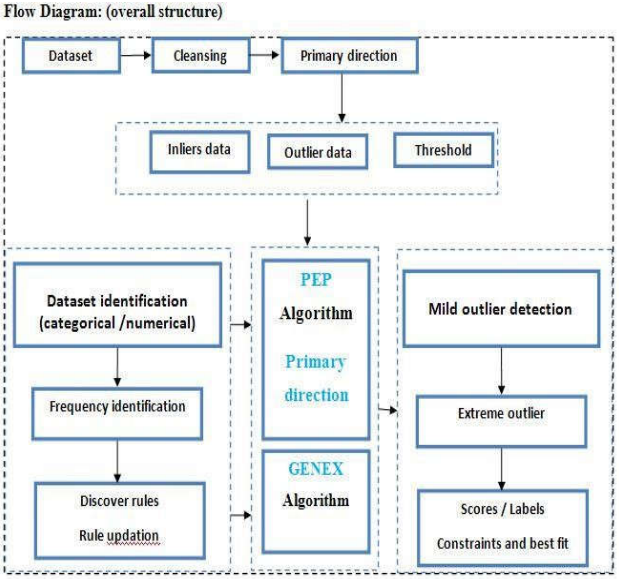

Fig 4.1: -Proposed Approach Framework

4.1. Methodology

The following algorithms and definitions help to identify the outlier effectively.

PEP - (Provisional Exceptional Property Pair extraction) algorithm: The PEP algorithm adopts a strategy consisting in selecting the relevant subsets of the overall set of conditions.

Step 1: Read dataset from high dimensional data
a) Read the attributes and values from the transaction TN.
b) Every attribute is set into a variable „a"
c) Set of condition is called „C"
Step 2: cleaning process Step 3: Pattern extraction
a) **Set C_a** as conditions -Identify base conditions for every attribute or properties
Step 4: Primary direction
a) Single clustered data set S_c
b) If the property is already in the cluster- find the label
c) Else if new attribute perform the following
d) Find next dimensionality
e) Find in next cluster
Step 5: set threshold-clustered data
Step 6: detect outlier from the dataset and return exceptional pairs
Step 7: rule updating process

Primary direction:
Step 1: Read every pattern
a) Check whether the dataset is numerical or categorical.
b) If numerical data then go to step 2
c) Else go to step 3.
Step 2: arrange the numerical dataset into ascending order, find median values and perform statistics modals.
Step 3: get categorical attributes and values. Verify the best fit of the selected pattern.
Step 4: returned dataset which is called as exceptional pair.
Step 5: return the extracted exceptional pair.

The above PEP algorithm for outlier detection uses a roll-up approach in which we test the outlier-like behavior of data points in different subspaces. The algorithm uses multi dimensional dataset as input parameters and the output pairs which define the outlier-like behavior of the data points. In addition, the maximum dimensionality r of the subspaces is input to the algorithm.

P in outlier detection
Recombination is the operation that copies individuals without modifying them. Usually, this operator is used to implement an elitist strategy that is adopted to keep the genetic code of the fittest individuals across the changes in the generations. If a good individual is found in earlier generations, it will not be lost during the evolutionary process.
The crossover operation allows genetic content exchange between two parents, in a process that can generate two or more children. In a GP evolutionary process, two parent trees are selected according to a matching (or pairing) policy and, then, a random tree is selected in each parent. Child trees are the result from the swap of the selected trees between the parents.
Finally, the **mutation operation** has the role of keeping a minimum diversity level of individuals in the population, thus avoiding premature convergence. Every solution tree resulting from the crossover operation has an equal chance of suffering a mutation process.

GENEX Algorithm: (GENetic Exception extraction)
Step 1: Generate base conditions:
a) Get outlier and inliers data"s and rules from the PEP modal
b) Threshold(maximum value)
c) If the data is categorical- add to the label
d) Else if a numerical attribute perform the following
✓ Based value it captures least data item of outlier and inlier.
✓ Repeat until checking complete all objects
Step 2: combine conditions.
Step 3: finally exprex algorithm generates exceptional pairs of data using genetic approaches by performing additional crossover and mutation process.
- Read set of outlier an inliers data attributes
- Declare a variable to store the output
- Set the condition (outlier dataset, inlierdataset, selected attribute, threshold outlier, threshold_inlier)
- For each property pairs set condition combination
- Combine conditions (outlier dataset, inlierdataset, selected attribute, conditions, threshold outlier, threshold_inlier)
- Proceed results

Comparison Graph

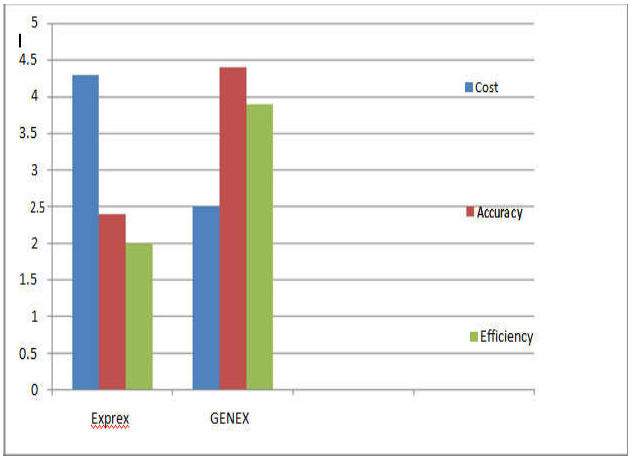

Fig 4.2: - Comparison Graph

5. Implementation

The implementation of the proposed system used visual studio environment with C#.net language. This chapter presents implementation and experiment which conducted by using the GENEX and PEP algorithm. The implementations are represents as follows. The consideration with some real data sets, including both numerical and categorical domains, in order to assess the capability of the approach in mining interesting knowledge. The implementation uses the initial direction which used initial principle analysis with the use of PEP algorithm. Then the system produces the graphical results which compute the frequent values and outliers among the dataset [7]. In order to point out differences and to show that the approach this presents a new and dynamic technique which is more powerful in characterizing groups of outliers effectively. This section also provides experimental results on both numerical and categorical datasets. This facilitates the implementation with the characterization and behavior analysis of the data by using the logical scenario.

6. Experiments

Steps to calculate Outlier for numerical dataset : Instructions

Sort the data in ascending order. For example take the data set {4, 5, 2, 3, 15, 3, 3, 5}. Sorted, the example data set is {2, 3, 3, 3, 4, 5, 5, 15}.
Find the median. This number at which half the data points are larger and half are smaller. If there are even numbers of data points, the middle two are averaged. For the example data set, the middle points are 3 and 4, so the median is (3 + 4) / 2 = 3.5.
Find the upper quartile, Q2; this is the data point at which 25 percent of the data are larger. If the data set is even, average the 2 points around the quartile. For the example data set, this is (5 + 5) / 2 = 5.
Find the lower quartile, Q1; this data point at which 25 percent of the data are smaller. If the data set is even, average the 2 points around the quartile. For the example data, (3 + 3) / 2 = 3.
Subtract the lower quartile from the higher quartile to get the interquartile range, IQ. For the example data set, Q2 - - Q1 = 5 -- 3 = 2.
Multiply the interquartile range by 1.5. Add this to the upper quartile and subtract it from the lower quartile. Any data point outside these values is a mild outlier. For the example set, 1.5 x 2 = 3. 3 -- 3 = 0 and 5 + 3 = 8. So any value less than 0 or greater than 8 would be a mild outlier. This means that 15 qualify as a mild outlier. Multiply the inter quartile range by 3. Add this to the upper quartile and subtract it from the lower quartile. Any data point outside these values is an extreme outlier. For the example set, 3 x 2 = 6. 3 -- 6 = -- 3 and 5 + 6 = 11. So any value less than --3 or greater than 11 would be a extreme outlier. This means that 15 qualify as an extreme outlier.

7. Conclusion

The proposed system implements a Genex algorithm which will create sub population for effective outlier detection. The implementation of LOO strategy with cross validation shows better result in the medical and numerical dataset. The system deeply verifies fitness value of the given transactional dataset with best and worst values. The Algorithm with various technology provided satisfactory results. Finding outliers and generating various sub population in

high dimensional dataset has been implemented. Finally the results shows the proposed system provides effective results for both numerical and categorical attribute, the output of the implementation is the score of mild and extreme outliers and exceptional property pair with labelling concept.
Using better algorithm and various techniques the system can be extended in the future. The future work may extend with effective un supervised technique and some fast computation techniques. The proposed system uses PEP algorithm with multiple clustering values. Re-clustering may be difficult when updating the closest values. The dynamic and random dataset may be used in future work.

REFERENCES

[1] C. C. Aggarwal, and P. S. Yu, Outlier detection for high dimensional data, ACM SIGMOD Conference on Management of Data, (2001).
[2] Ji Zhang, Meng Lou, Tok Wang Ling and Hai Wang, HOS-Miner: A System for Detecting Outlying Subspaces of High-dimensional Data, In: Proc. Int"l Conf. Very Large Databases (VLDB "04), Toronto Canada, 2004.
[3] Ji Zhang, Qiang Gao and Hai Wang, A Novel Method for Detecting Outlying Subspaces in Highdimensional Databases Using Genetic Algorithm, In: Proc. Int"l Conf. Data Mining (ICDM "06), 2006.
[4] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In Proc. 24th VLDB, pages 392–403, 24–27 1998
[5] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. pages 427–438, 2000.
[6] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
[7] I. T. Jolliffe. Principal Component Analysis. Springer Verlag-New York, 2nd edition, 2002.
[8] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artificial Intelligence Rev., vol. 22, no. 2, pp. 85 - 126, 2004.
[9] N.V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," SIGKDD Explorations, vol. 6, no. 1, pp. 1-6, 2004.
[10] E. Knorr and R. Ng, "Algorithms for Mining Distance-Based Outliers in Large Data Sets," Proc. Int"l Conf. Very Large Data Bases (VLDB" 98), 392-403, 1998.
[11] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High- Dimensional Data Sets," IEEE Trans. Knowledge and Data Eng.,vol. 17, no. 2, pp. 203 -215, Feb. 2005.
[12] F. Angiulli and F. Fassetti, "Dolphin: An Efficient Algorithm for Mining Distance-Based Outliers in Very Large Data Sets," ACM Trans. Knowledge Discovery from Data, vol. 3, no. 1, article 4, Mar. 2009.
[13] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying Density-Based Local Outliers," Proc. ACM SIGMOD Int"l Conf. Management of Data (SIGMOD), pp. 93-104, 2000,
[14] Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, *9*(1), 307.
[15] Tiba, A., Bartik, Z., Toman, H., & Hajdu, A. (2019, September). Detecting outlier and poor quality medical images with an ensemble-based deep learning system. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)* (pp. 99-104). IEEE.
[16] Lile, C., & Yiqun, L. (2017, September). Anomaly detection in thermal images using deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 2299-2303). IEEE.
[17] Isaksson, C., & Dunham, M. H. (2009, July). A comparative study of outlier detection algorithms. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 440-453). Springer, Berlin, Heidelberg.
[18] Viswanath, S., & Madabhushi, A. (2012). Consensus embedding: theory, algorithms and application to segmentation and classification of biomedical data. *BMC bioinformatics*, *13*(1), 26.
[19] Xu, X., Liu, H., Li, L., & Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, *11*(1), 652-662.