WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A NOVEL GENETIC APPROACH FOR OUTLIER DETECTION

Mr. M. Madhangiri, Mrs. S. Mekala

Department of Business Administration, Park's College, Tirupur

ABSTRACT:

A noisy data elimination, attribute and property discovery is a major consideration in the proposed method. From the overall given population the system predicts the sub population effectively. The subpopulation and exceptional property pair which is known as outliers. With the aim of effective outlier detection, the proposed PEP algorithm applies a provisional model which identifies the exceptional property par with the best fit method implementation. There are several outlier detection methods have been introduced with certain domains and applications, but the techniques were more generic and suffer from confidentiality problem. The proposed concept effectively implements Genetic modal based approach which is named as GENEX algorithm and PEP algorithm for the detection of sub population scores for both numerical and categorical datasets. Additionally the system performs the best fit method in order to find best class based on the score and label. The proposed algorithm can reduce the computation cost and lack of accuracy problem by applying best data mining and suitable pruning techniques. The experiments and the results provide the mild and extreme outlier ranges with best fit values.

Keywords: Genetic Approach, Outlier Detection, PEP algorithm.

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.

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

2. APPLICATIONS OF OUTLIER DETECTION

Outlier detection has been a widely researched problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas.

Finding subpopulation in outlier is a tedious process; existing system only identifies the abnormal behavior as outlier instead of range calculation.

2.1. Existing system:

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.
The existing system discovers attributes or properties based on the given populations which are called as inliers. Existing methods are,

- **Probabilistic Model**
- **SVM based outlier Model**
- **Statistical modal**
- **Distance based approaches**

2.2. 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.

3. PROPOSED SYSTEM

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. 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.

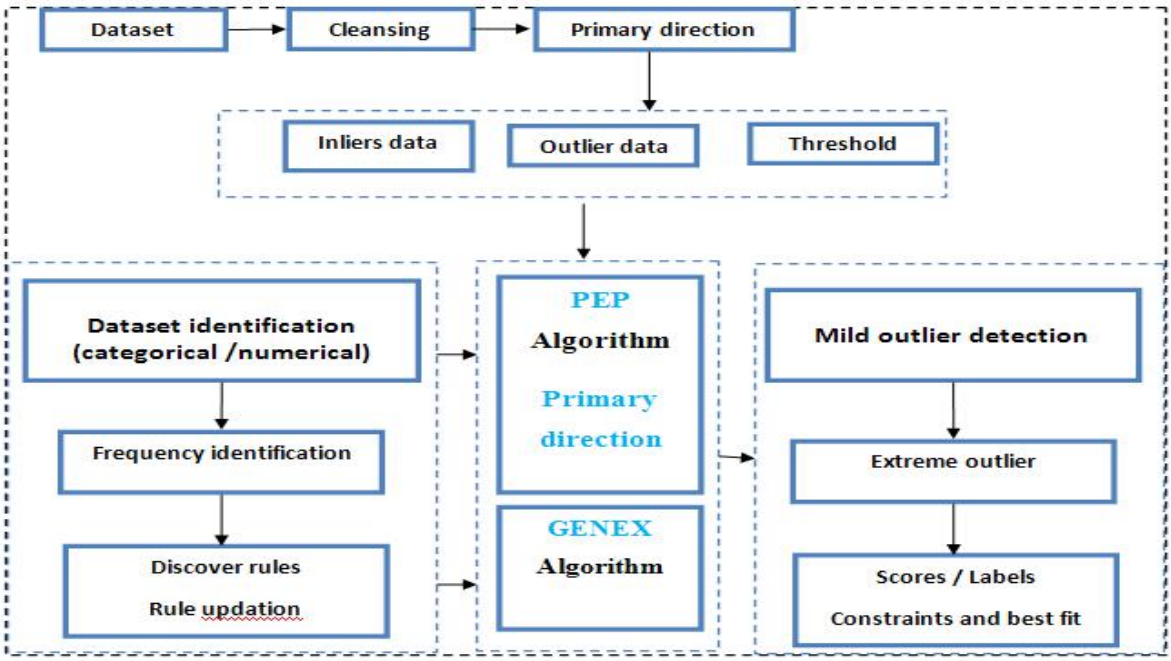
Moreover, each property can have associated a condition, also called explanation, whose aim is to single out a (significant) portion of the data for which the property is indeed characterizing anomalous subpopulations. In order to single out significant properties, this resorts to minimum distance estimation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

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.

Flow Diagram: (overall structure)



4. 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.

GP 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.

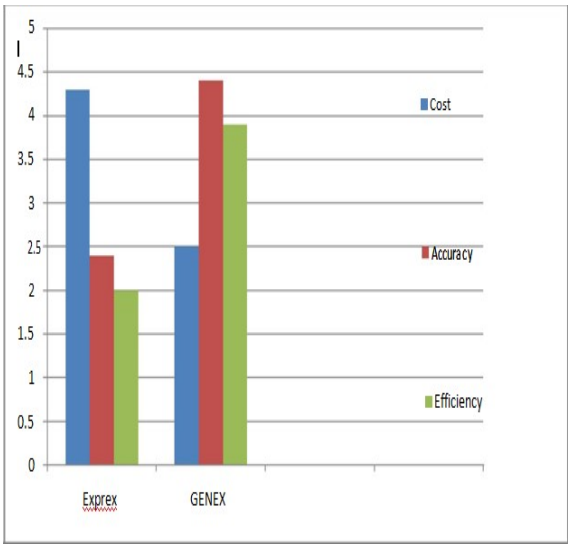
Step3: finally exprex algorithm generates exceptional pairs of data using genetic approaches by performing additional crossover and mutation process

- 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

5. PERFORMANCE EVALUATION

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. 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.



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.

6. CONCLUSIONS AND FUTURE WORK

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 labeling 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] "Outlier Detection Algorithms in Data Mining" Jingke Xi ; Sch. of Comput. Sci. & Technol., China Univ. of Min. & Technol., Xuzhou.

[2] "Association rules based algorithm for identifying outlier transactions in data stream" Li-Jen Kao ; Yo-Ping Huang Systems, Man, and Cybernetics (SMC), 2012 IEEE

[3] A Survey of Outlier Detection Methods in Network Anomaly Identification Prasanta Gogoi1" , D K Bhattacharyya1 , B Borah1 and Jugal K Kalita2

[4] C. C. Aggarwal, and P. S. Yu, Outlier detection for high dimensional data, ACM SIGMOD Conference on Management of Data, (2001).

[5] 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.

[6] 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.
[7] Statistical Methods for Research Workers. Oliver and Boyd, 1954.
[8] J. Kubica and A. Moore. Probabilistic noise identification and data cleaning, 2002.
[9] S. Schwarm and S. Wolfman. Cleaning data with bayesian methods, 2000
[10] 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
[11] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. pages 427–438, 2000.
[12] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
[13] I. T. Jolliffe. Principal Component Analysis. Springer Verlag-New York, 2nd edition, 2002.
[14] V. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Rev., vol. 22, no. 2, pp. 85- 126, 2004.
[15] 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.
[16] 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), pp. 392-403, 1998.
[17] 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.
[18] 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.
[19] 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,
[20] S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, “LOCI: Fast Outlier Detection Using the Local Correlation Integral,” Proc. Int‟l Conf. Data Eng. (ICDE), pp. 315-326, 2003,
[21] F. Angiulli, G. Greco, and L. Palopoli, “Outlier Detection by Logic Programming,” ACM Trans. Computational Logic, vol. 9, no. 1, article 7, 2007.
[22] F. Angiulli, R. Ben-Eliyahu-Zohary, and L. Palopoli, “Outlier Detection Using Default Reasoning,” Artificial Intelligence, vol. 172, nos. 16/17, 1837-1872, Nov. 2008.