A Suggested method for detecting outliers based on a particle swarm optimization algorithm

Emad Obaid Merza¹ and Nashaat Jasim AL-Anber²

¹ Information Technology Department, Technical College of Management- Baghdad
Middle Technical University, Baghdad, Iraq. Email: edmistry1974@gmail.com

² Information Technology Department, Technical College of Management- Baghdad
Middle Technical University, Baghdad, Iraq. Email: dr.nashaat@mtu.edu.iq

Abstract

The occurrence of tremendous developments in the field of data has led to the formation of huge volumes of data, and it is natural that this leads to the presence of outliers in this data for many reasons. The presence of outliers in the data affects the statistical analysis, so we must try to reduce their impact in various ways. On the other hand, the existence of outliers may be of great benefit in many application and of great importance in various fields. In this paper we propose a new method for detecting outliers based on the Particle Swarm Optimization algorithm (PSO). The new propose algorithm was compared with the normal distribution method, and the results obtained from the new method were very promising and encouraging.

1. Introduction

The importance of analyzing outliers increases with the acceleration of development and broad jumps in information technology, as data volumes have become more inflated and complex, which requires converting these data into useful information for the decision-making process and data analysis, and that transformation process includes in its context a very important matter, which is the concept of outliers [1]. The issue of outliers has been taken up by many scientists and researchers in order to study the effect of these values on the accuracy of the results that expected from the data analysis process and among those prominent scientists who have dealt with the concept of outliers is Hawkins and Freeman. The outliers in a particular data set may appear in the form of one or more values, what
distinguishes this value is that it is not logical in relation to the rest of the natural data, for example, it may be very large or very small compared to the mean of the data [1,2], and that the existence of a unique value is of high importance, because it has important implications in data mining as well as in analyzing medical and financial data and in the field of networks, as detection of intrusion on networks is one of the most applied topics that have gained importance in recent years [3]. The exploration of outliers or unique patterns is a very important sub-topic in data mining, as the process of detecting outliers is the exploration of unique patterns that clearly deviate from the natural path of the data, the interest in classifying the data and knowing their behavior and common characteristics and differences between them. It will inevitably lead to ease of study [1], and the concept of outliers is always associated with the study of the natural behavior of the data, as any behavior that does not resemble normal behavior is considered an outlier behavior, For example, unnatural geological activities that precede natural disasters. The process of detecting outliers is applied in many sectors, for example in the health sector, as it is used to diagnose diseases, the financial sector is used to detect fraud in credit cards, and in networks, the abnormal activity in the network may mean the presence of an infiltration, attack or entry process unauthorized and outliers is used in image processing as well as in the industrial sector to detect industrial defects in products. In statistics, the idea of outliers can be discussed on the basis that they do not share characteristics with the community or the sample. The community is defined as a group of beings who have common characteristics and the sample is part of the community and bears all of its characteristics [1]. Sometimes the existence of outliers may not be meaningful and the first important process in statistical analysis of data is to identify and detect outliers because their presence may lead to misleading results [4]. There are a large number of techniques for detecting outliers that can be summarized by statistical methods, Density-Based Methods, Cluster-Based Methods, Distance-Based Methods, Subspace-Based Methods and Deviation-Based Methods. Multiple outliers are detected one by one in the regression analysis model, and this may lead to misleading results due to the smearing and masking effect, and finding techniques that detect outliers at once will avoid these effects [5]. Detecting outliers is important because it affects the estimated model of the regression, especially when using the least squares method. Therefore, detection of outliers is very important in practical applications [6], since presence of outliers resulting from entering a value by mistake or from a different distribution will affect the slope of the regression line, as the outliers, higher or lower, serve to push the mean of data away from its position, which leads to raising or lowering the regression line [7]. Evolutionary algorithms have been used to transform the problem of detecting outliers into an improvement problem, as the Particle Swarm Optimization (PSO) algorithm has been used to improve the automatic
calculation of distances between data points instead of normal statistical methods that can have potentially error, and the use of the genetic algorithm gave great results in detecting outliers [8,9]. And many previous studies have used evolutionary algorithms to improve detection of outliers are as follows:

Crawford and Wainwright (1995) developed the Genetic Algorithms (GA) to manage the process of selecting a sub-sample of potentially extreme data from a general data population. Each data point in the sub-sample has its own list of indicators and the sub-sample will be considered an ordered gene. The point markers on the far left of the gene are outliers and the rest are normal. This research demonstrated that GA performed excellent in detecting multiple outliers [8]. Akter and Khan (2010) used an evolutionary quantum-inspired algorithm (QEA) to detect multiple outliers at once in a linear regression analysis model. It was found that QEA overcame smearing and masking effects that lead to misleading results when using regular statistical methods that expose outliers one by one [5]. In the same year, Mohemmed and Brawne used the Particle Swarm Optimization (PSO) algorithm to improve measures of automatically calculating the distance between data points to detect outliers and compare the results of this method with Local Outlier Factor (LOF) after applying it to five sets of real data. The results showed the superiority of the PSO method over the LOF method [9]. And too in the same year, Wahid and Rao used Particle Swarm Optimization (PSO) to detect outliers in high-dimensional data by calculating the degree of anomaly based on calculating the distance between the data point and its closest neighbors, and then applying the PSO to the degree of anomaly according to a pre-set threshold value and comparing the results of this method with the results of High Contrast Subspaces for Density-based Outlier Ranking (Hics) algorithm, the two methods were tested on data from UCI repository and the PSO method was the most accurate and efficient of Hics [10]. Afzal and Ashraf (2016) used the genetic algorithm (GA) to detect outliers. The researchers indicated that the reason for using GA is its adaptability and that the results using the GA algorithm were encouraging and efficient even with the increase in data sets. The type of data used in that research are breast cancer and diabetes medical data and commercial data for a market [11]. Alguliyev et al. (2019) working on improving the accuracy of detection of outliers by the k-mean algorithm by proposing a weighted grouping method, which is a combination between Particle Swarm Optimization (PSO) and k-mean algorithm. A test of the new method was performed on S5 files on Yahoo site, and the test result indicated that the new method is more accurate than the k-mean algorithm [12].
2. Outliers

Outlier, as defined by Hawkins as a value that is distinctly different from other values, which raises the suspicion that it was generated in a way that differs from the rest of the normal values, and it is simply a value that differs from the rest of the other values [1], and statistically, the normal values are obtained by means of a mechanism. Accordingly, the outliers are the ones that deviate significantly from this generation mechanism [3,2], that is, the outliers originate from different generation sources [1] that is, the value is called an outlier if it is not similar to most other objects in terms of the predominant characteristics. for data [13], as shown in Figure(1).

![Figure (1): Shows the outliers represented by points p, s and group c](image1)

There are many types of outliers, as the dataset can contain multiple types of outliers [3] which are **Global outliers** Figure(2), **Context outliers**, **Collective outliers** Figure(3).

![Figure (2): The values in region D represent a Global outliers](image2)

![Figure (3): The points in bold color represent collective outliers](image3)
The outliers are appear for many reasons, for example: distributions that differ from the distribution that produced the normal values in the dataset, measurement error, data entry errors, data is generated using a mixture of different distributions types and error in processing the data [1,3,2]. There are several common methods for detecting outliers that can be categorized into several methods: (item labels based methods, statistic based methods, proximity-based methods and cluster-based methods).

3. Particle Swarm Optimization algorithm (PSO)

PSO is one of the evolutionary algorithms that simulate the intelligence of swarms in nature, such as flocks of birds and fish, and the precise idea of it is based on the fact that the elements of the swarm have individual solutions on their own and move irregularly in the search area in search of the best location for each element and then the best location among the element sites, as each element of the swarm changes its direction of movement to reach the best location, and then the new location for each element is calculated after calculating the new velocity for each element, which is the sum of the velocity and the old location of the element. In other words, PSO idea is that the particle swarm is stationed in random locations and moves randomly within the search range and the direction of the particles changes in search of new sites better than the previous sites and to find the best location for each particle of the Best Position (BP), this is done by finding the new velocity. \( V_{i}(t + 1) \) and according to equation (1), then the new position (BP) is calculated for each particle of the swarm according to Equation (2). For the purpose of finding the best location for all the particles of the swarm, the Global Best Position (GB) requires a Fitness Function which It is calculated for each particle of the swarm, and through it we can know the global best GB position for all swarm elements since the global best position corresponds to the highest value of the fitness function [14] [15].

\[
V_{i}(t + 1) = WV_{i}(t) + C_{1}r_{1}(p_{i}(t) - x_{i}(t)) + C_{2}r_{2}(GB - Xi(t))
\]  \hspace{1cm} (1)

\[
X_{i}(t + 1) = X_{i}(t) + V_{i}(t + 1)
\]  \hspace{1cm} (2)

where

Velocity \( V_{i} \) represents the velocity of each squadron element

New Velocity \( V_{i}(t + 1) \)

Position \( X_{i} \) Represents the location of each swarm element
New **Position** $X_i(t + 1)$

**Random Variables** $(r_1, r_2)$ represent randomly generated numbers within a period $[0, 1]$

**Constants** $(C_1, C_2)$ represent the acceleration parameters and it should be $(C_1 + C_2 = 4)$

**Inertia weight** $(W)$ is the inertia weight.

The PSO algorithm is one of the easiest evolutionary algorithms in the application and by using it to solve problems of outlier detection, we convert the problem of detection of outliers into an optimization problem, and this will lead to the detection method being more sensitive to outliers, which leads to higher accuracy in detecting outliers [9]. In our suggested method we'll pause the implementation of the PSO algorithm at the first convergence which occurs in the second iteration of the PSO algorithm implementation.

4. **Proposed method:**

**Control Chart Technique using Linear Regression Based on Particle Swarm Optimization Algorithm (CCT-LR-PSO)**

This proposed method aims to achieve the most accurate results possible in the process of detecting outliers as well as setting more realistic limits for the normal data by developing a Control Chart Technique (CCT) [16] using the regression line instead of the mean of data, and then using one of the evolutionary algorithms which is the particle swarm optimization algorithm (PSO), although the PSO algorithm is suitable for a specific number of data points, as some studies mention that this number may reach 60 data points, but the part of the (CCT) using the regression line had a positive effect on increasing this number and making this hybrid technology exceed the issue of its suitability for a specific number. It can be applied to data points higher than 60 points, although errors were recorded in recording data points as outliers and they are not. However, the accuracy recorded for this method proved its superiority over the normal distribution method within the limits of $(\mu \pm 3\sigma)$ and the results of this method were very promising in the reports that It was done in order to verify its accuracy (the tests will be discussed in detail later in Section5), and the researcher believes that despite the use of the PSO algorithm, the method did not go to more complexity and preserved It is easy to use in a manner consistent with the methodology and vision of this research,
although the combination of accuracy and ease of use in any method for
detecting outliers is not an easy matter, as it is accepted that investigating
accuracy in a specific work leads to an increase in calculations or
consumption of time or resources generally. The following are the steps for
applying the (CCT-LR-PSO) method:

1. Find a linear regression equation for the data.
   \[ y = \beta_0 + \beta_1 x \quad (3) \]

2. Apply the regression equation to all x-axis points from X1 to Xn, The
   results of this process will represent the particle swarm (Pai).

3. Implementation of the PSO swarm optimization algorithm on all swarm
   elements to the second iteration to find (MaxPai) and (MinPai) and to
   maximize the two fitness function values as follows:
   \[
   \text{max fitness function} = \text{MaxPai} + \mu + \left(\frac{\mu}{0.25 \times n}\right) \quad (4)
   \]
   It is used to find the highest value for normal data

   \[
   \text{max fitness function} = \text{MinPai} - \mu - \left(\frac{\mu}{0.25 \times n}\right) \quad (5)
   \]
   It is used to find the minimum value for normal data

   where MaxPai represents the largest element of the swarm and MinPai
   represents the smallest element of the swarm, \(\mu\) is the arithmetic mean of
   the data \(n\) the number of data points.

4. To find the highest outlier, and after implementing the PSO algorithm to
   the second iteration, depending on equation (4) as a fitness function, and
   any value will be considered as an outlier if it is greater than the lowest
   value of the fitness function in the second iteration.

5. To find the minimum outlier and after implementing the PSO algorithm
   to the second iteration, depending on equation (5) as a fitness function,
   then any value will be considered as an outlier if it is less than the
   lowest value of the fitness function in the second iteration.

6. It is worth noting that if \(\beta_1\) is negative, then we apply the regression
   equation at \(X = 1\) to find the value of MaxPai and at \(X = n\) to find
   MinPai directly and apply equations (4) and (5) to find the upper and the
   lower values of the data, respectively.
Below Figure No. (4) represents the general structure of the (CCT-LR-PSO) algorithm.

Figure No.(4) represents the general structure of the CCT-LR-PSO method.
5. The Application

Synthesized data (normal data contaminated with fake outliers) was used to test the proposed method and compare it with the normal distribution method, as well as the production lines data of the Abu Ghraib dairy factory for the year 2019 for the products (cream, cheddar cheese, yogurt, Shanina milk, free fat, butter, cheese).

5.1 experimental data

The researcher used data distributed normally within \((\mu + 3\sigma)\) and added to it fake outliers in order to test and compare the proposed method with the normal distribution method, the tested data are the data of the months \((2, 3, 4, 8, 10, 12)\) for the cream production line and the reason for choosing these months because both methods did not detect any outlier in these months, as well as the two methods were tested on data normally distributed within \((\mu + 3\sigma)\) and with sizes \((50, 100, 200)\) and Table No. (1) shows the original used sample sizes and the number of added outliers.

| Sample size | Size after adding | Added outliers | Notes                  |
|-------------|-------------------|----------------|------------------------|
| 18          | 20                | 2              | February Data          |
| 10          | 11                | 1              | March Data             |
| 8           | 9                 | 1              | April data             |
| 19          | 21                | 2              | July data              |
| 23          | 24                | 1              | October data           |
| 27          | 31                | 4              | December data          |
| 50          | 55                | 5              | Synthesized data       |
| 100         | 107               | 7              | Synthesized data       |
| 200         | 210               | 10             | Synthesized data       |

5.2 Compare CCT-LR-PSO Method with Normal Distribution Method

The proposed CCT-LR-PSO method and the normal distribution method were applied to the data. After adding the dummy outliers, the (CCT-LR-PSO) method proved to be superior in all samples although it missed normal values as being outliers in the data sizes \((107 \text{ and } 210)\) since the normal distribution method did not reveal any outlier in all the samples that were tested, and Table (2) Results of the detection test for outliers using the normal distribution method and the CCT-LR-PSO method using the upper limit of the data according to the normal distribution method.
Table (2): Shows the test results on the two methods according to the upper limit of the normal distribution method

| Sample size | Size after adding | Number of detected outliers | normal distribution | CCT-LR-PSO |
|-------------|-------------------|-----------------------------|---------------------|------------|
|             |                   |                             |                     |            |
| 18          | 20                | 0                           | 2                   |            |
| 10          | 11                | 0                           | 1                   |            |
| 8           | 9                 | 0                           | 1                   |            |
| 19          | 21                | 0                           | 2                   |            |
| 23          | 24                | 0                           | 1                   |            |
| 27          | 31                | 0                           | 4                   |            |
| 50          | 55                | 0                           | 5                   |            |
| 100         | 107               | 0                           | 11                  |            |
| 200         | 210               | 0                           | 12                  |            |

The accuracy scale derived from the confusion matrix was used to evaluate the results of the two methods, which appeared in Table (2), as follows:

\[
Accuracy = \frac{TP + TN}{P + N}
\]

\[
\text{ACCURACY}_{\text{CCT-LR-PSO}} = \frac{39 + 455}{500} = 0.988
\]

\[
\text{ACCURACY}_{\text{ND}} = \frac{0 + 455}{494} = 0.921
\]

In order to ensure the superiority of the CCT-LR-PSO method, the researcher used the upper limits of the natural data for the CCT-LR-PSO method instead of the upper limits of the normal data for the normal distribution method to add the same number of fictitious outliers and re-test the two methods and the results were as shown in Table (3).

Table (3): Shows the new test results on the two methods according to the upper limit of CCT-LR-PSO method

| Sample size | Size after adding | Number of detected outliers | normal distribution | CCT-LR-PSO |
|-------------|-------------------|-----------------------------|---------------------|------------|
|             |                   |                             |                     |            |
| 18          | 20                | 0                           | 2                   |            |
| 10          | 11                | 0                           | 0                   |            |
| 8           | 9                 | 0                           | 1                   |            |
| 19          | 21                | 0                           | 1                   |            |
| 23          | 24                | 0                           | 1                   |            |
| 27          | 31                | 0                           | 0                   |            |
| 50          | 55                | 0                           | 6                   |            |
| 100         | 107               | 0                           | 11                  |            |
| 200         | 210               | 0                           | 12                  |            |
In order to measure the accuracy of both methods according to this new test, we use the same accuracy measurement method that was used previously, and the measurement results were as follows:

\[
{\text{ACCURACY}}_{\text{CCT-LR-PSO}} = \frac{34 + 455}{501} = 0.976
\]

\[
{\text{ACCURACY}}_{\text{ND}} = \frac{0 + 455}{494} = 0.921
\]

It is mentioned in the statistical literature of the field of detection of outliers that building a comprehensive model for the detection of outliers is an impossible task. The researcher believes that this statement is completely correct, as the degree of perfection cannot be reached in any field of scientific research, but if the work is directed towards normal limits for more realistic data, the results will be very close to perfection and to high degrees of accuracy. Therefore, the proposed method was more sensitive to outliers than the normal distribution method because its effectiveness lies in that it defined realistic limits for the area and limits of normal data. The figures (5), (6), (7) shows the graph of data flocks of size (50, 100, 200) respectively depending on the higher boundaries of the data, once for the normal distribution method and again for the CCT-LR-PSO method. It is evident that the CCT-LR-PSO method reduced the limits of the normal data presence to realistic limits, which led to the possibility of detecting the added outliers, while the normal distribution method failed to detect the added outliers.

Figure (5) Section (a) represents the addition of outliers using the upper limits of the normal data for the normal distribution method and the section (b) represents the addition of outliers using the upper limits of the normal data for the CCT-LR-PSO data for size (50 data points)
Figure (6) Section (a) represents the addition of outliers using the upper limits of the normal data for the normal distribution method and the section (b) represents the addition of outliers using the upper limits of the normal data for the CCT-LR-PSO data for size (100 data points).

Figure (7) Section (a) represents the addition of outliers using the upper limits of the normal data for the normal distribution method and the section (b) represents the addition of outliers using the upper limits of the normal data for the CCT-LR-PSO data for size (50 data points).

Below we include the results of applying the proposed CCT-LR-PSO method and the normal distribution method within $(\mu + 3\sigma)$ to the data of the Abu Ghraib factory, as Table (4) shows the results of applying the two methods to the production data day / year (229 production days). The sample size represents for all production lines, and Table No. (5) shows the results of applying the two methods to the production data for
a month / year (12 months represents the sample size) for all production lines, and Table No. (6) shows the results of applying the two methods to the data of the curd production line on a day /Month

Table (4): Shows the number of outliers detected using the two methods on production data day/year and by product

| product name          | Number of detected outliers |
|-----------------------|-----------------------------|
|                       | Normal Distribution | CCT-LR-PSO |
| Cream                 | 5                          | 41         |
| Cheddar cheese        | 1                          | 25         |
| Yoghurt               | 4                          | 22         |
| Laban Shanina         | 9                          | 29         |
| Butter                | 8                          | 34         |
| Free fat              | 10                         | 22         |
| Soft cheese           | 11                         | 16         |

Table (5): Shows the number of outliers detected using the two methods on production data month/year and by product

| product name          | Number of detected outliers |
|-----------------------|-----------------------------|
|                       | Normal Distribution | CCT-LR-PSO |
| Cream                 | 0                          | 2          |
| Cheddar cheese        | 0                          | 1          |
| Yoghurt               | 0                          | 1          |
| Laban Shanina         | 0                          | 1          |
| Butter                | 0                          | 2          |
| Free fat              | 0                          | 1          |
| Soft cheese           | 0                          | 3          |
Table (6): represents the outliers detected in the cream product data that represent daily / month production for the year 2019 using the two methods

| Month      | Sample size | Number of detected outliers | Normal Distribution | CCT-LR-PSO |
|------------|-------------|-----------------------------|---------------------|------------|
| January    | 20          | 2                           | 0                   | 2          |
| February   | 18          | 0                           | 0                   | 0          |
| March      | 10          | 0                           | 0                   | 0          |
| April      | 8           | 0                           | 0                   | 0          |
| May        | 20          | 0                           | 0                   | 1          |
| June       | 15          | 0                           | 2                   |            |
| July       | 22          | 0                           | 1                   |            |
| August     | 19          | 0                           | 0                   |            |
| September  | 24          | 0                           | 0                   |            |
| October    | 23          | 0                           | 0                   |            |
| November   | 23          | 0                           | 2                   |            |
| December   | 27          | 0                           | 0                   |            |

6. Results

The tests proved the superiority of the proposed method over the normal distribution method, and it is worth noting that the normal distribution method did not detect any outlier when testing the two methods according to the upper boundaries of the data according to the CCT-LR-PSO method, although the values of these limits are less than the values of the upper limits of the data according to the method normal distribution, and the researcher believes that the reason is due to the loose limits of normal data that are adopted by the normal distribution method, so devising methods that lead to making the normal data boundaries more realistic is what future work should be in and that suits the nature of each type of data.
7 References

[1] Aggarwal, C. C. (2016). Outlier analysis. Springer-Verlag.

[2] Kriegel, H. P., Kröger, P., & Zimek, A. (2010). Outlier detection techniques. Tutorial at KDD, 10, 1-76.

[3] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 83-124.

[4] Breaban, M., & Luchian, H. (2012). Outlier detection with nonlinear projection pursuit. International Journal of Computers Communications & Control, 8(1), 30-36.

[5] Akter, S., & Khan, M. H. (2010). Multiple-Case Outlier Detection in Multiple Linear Regression Model Using Quantum-Inspired Evolutionary Algorithm. JCP, 5(12), 1779-1788.

[6] Alma, O. G. (2013). Performances Comparison of Information Criteria for Outlier Detection in Multiple Regression Models Having Multicollinearity Problems using Genetic Algorithms. MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics, 29, 119-131.

[7] Rousseeuw, P. J., & Leroy, A. M. (2005). Robust regression and outlier detection (Vol. 589). John wiley & sons.

[8] Crawford, K. D., & Wainwright, R. L. (1995, July). Applying Genetic Algorithms to Outlier Detection. In ICGA (pp. 546-550).

[9] Mohemmed, A. W., Zhang, M., & Browne, W. N. (2010, July). Particle swarm optimization for outlier detection. In Proceedings of the 12th annual conference on Genetic and evolutionary computation (pp. 83-84).

[10] Wahid, A., & Rao, A. C. S. (2019). A distance-based outlier detection using particle swarm optimization technique. In Information and Communication Technology for Competitive Strategies (pp. 633-643). Springer, Singapore.
[11] Afžal, M., & Ashraf, S. A. (2016). Genetic algorithm for outlier detection. Int. J. Comput. Sci. Inf. Technol.(IJCSIT), 7(2), 833-835.

[12] Alguliyyev, R. M., Alguliyyev, R. M., & Abdullayeva, F. J. (2019). PSO+ K-means algorithm for anomaly detection in Big Data. Statistics, Optimization & Information Computing, 7(2), 348-359.

[13] SURI, N. M. R., Murty, M. N., & Athithan, G. (2019). Outlier Detection: Techniques and Applications. Springer Nature.

[14] Talukder, S. (2011). Mathematical modelling and applications of particle swarm optimization.

[15] Tharwat, A., Gaber, T., Hassanien, A. E., & Elnaghi, B. E. (2017). Particle swarm optimization: a tutorial. In Handbook of research on machine learning innovations and trends (pp. 614-635). IGI global.

[16] SkyMark: Control Chart, at http://www.skymark.com/resources/tools/control_charts.asp (accessed: 23 August 2020)