

Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm

A.R. Ajiboye, A.G. Akintola & A.O. Ameen

Department of Computer Science
University of Ilorin
Ilorin, Nigeria

ajibraheem@live.com, abimbolar@yahoo.com, aminamed@unilorin.edu.ng

J. Abdul-Hadi

Department of Mathematics
Bauch State University
Gadau, Nigeria
jamcy98@gmail.com

ABSTRACT

The purity of the dataset used for model construction plays important roles in the accuracy and reliability of model building; outliers are often caused by noisy data as a result of mechanical faults, changes in system behaviour, or due to human error. This is why it is essential to pre-process dataset prior to modelling, in order to differentiate between data that appears normal or abnormal within the sample space. One important reason for removing outliers is to prevent contaminating effect on the dataset which can lead to bad consequences and serious disaster if not removed. An effective measure that automatically clusters outliers in the dataset using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique is proposed in this paper. Rapidminer, an open source software tool is used to experiment on some sample dataset and based on the characteristics of these data objects, some clusters are formed which filter out outliers from the dataset being explored. The experimental results from this study show that, the DBSCAN algorithm is a suitable technique for outliers detection and capable of filtering the abnormal data from a combination of noise and normal dataset.

Keywords: Anomaly detection, DBSCAN, clustering, model-building, algorithm, noisy data.

African Journal of Computing & ICT Reference Format:

A.R. Ajiboye, J. Abdul-Hadi, A.G. Akintola & A.O. Ameen (2015). Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm. Afr J. of Comp & ICTs. Vol 8, No. 1. Pp39-46.

1. INTRODUCTION

A dataset is said to consist of anomalies if it shows some patterns that do not conform to a well-defined notion that depicts normal behaviour. This may be as a result of the presence of noise. Noise is a random error or variance in a measured variable [1]. A serious problem with any form of automatic detection of apparently incorrect data is the fear of removing legitimate data along with the abnormal data. One-class classification is often used to differentiate between data that appears normal and abnormal with respect to the distribution of the training data [2]. The generic statistical approach to one-class classification is to identify outliers as instance that lie beyond a distance d from a given percentage p of the training data.

Alternatively, estimating the probability density for the target class by fitting a statistical distribution, such as Gaussian, to the training data; any test instances with a low probability value can be marked as outliers [2], however, it is a big challenge to identify an appropriate distribution for the data. Three approaches to anomaly detection are fundamental: The detection of outliers without having prior knowledge of the data. This is synonymous to a learning approach using unsupervised technique such as clustering; the second approach is modelling both normal and abnormal data and using a pre-labelled data as target. This is supervised learning such as in classification problem. The third approach is to model only normal data or in some few instances, model the abnormal data [3], this is analogous to a semi-supervised recognition or detection task, authors generally referred to this approach as novelty detection or novelty recognition [4].

The removal of Outlier detection is a critical task in many safety environments as the presence of outlier is a clear sign of abnormal running conditions from which significant performance degradation may be the consequence [4]. The concept of Semi-Discrete Decomposition (SDD), can be used to decide which objects or values are outliers, as SDD is a transformation in a geometric space [5]. For instance, nodes in the hierarchical clustering that contain one, or just a few objects might be considered to describe outliers. However, in a situation where the hierarchical clustering extend far, then, every object is in a node by itself, therefore, it is only the nodes with small number of object near the root of the tree that are really outliers [5].

In using the clustering technique to detect outliers, where similar values are organized into group or clusters, intuitively, values that are not within the set of clusters may be considered outliers [1]. In a statistical regression, outliers is usually obvious if, for instance, a straight line is being fitted to data that lies on a bell curve, however, most problems cannot be easily visualised. A good solution can be found by applying several different learning techniques to filter the data and anomalies detected by the filter can then be examined by a human expert [6]. Other way to characterize outliers in a dataset is their participation in many bumps as such objects have attribute values that overlap with other common attribute values but usually few in number [5]. Also, efforts are made at removing outliers in [7], the study minimizes the cost function using standard optimization techniques; the study also suggested that, if the residual error is too great, the offending measurement can be removed and optimization can continue.

In data mining applications, such as clustering, outlier analysis, and nearest neighbour classification, these techniques helps to make comparison of objects to determine the degree of their similarities. There are several ways by which the concept of outliers detection can be put to use. The concept is applicable in areas such as: fraud detection, discovery of the intrusion to a network, video surveillance, medicine, pharmaceutical test, weather forecast etc.

The notion that constitutes anomalies has been observed as what differs for different application domains. Chandola et al. [8] illustrates instances in the medical domain where a small deviation from normal such as little rise in body temperature might be an anomaly, while, similar deviation in the stock market domain such as fluctuation in the value of stock might be seen as normal. Thus, applying a technique developed in one domain to another is not usually straightforward.

The clustering approach that is based on density-based distribution of data is proposed in this study to detect the outliers in a dataset with a view to improving model accuracy and reliability. The rest of the paper is organized as follows: In the next section, the dense-region modelling of clusters is discussed. In section 3, some works related to this study is reviewed, while in section 4, the material and methods used for the study is discussed. The experimental results is displayed and discussed in section 5 and the paper is concluded in section 6.

2. DENSE-REGION CLUSTERS MODELLING

The strategy behind density-based clustering method is to model clusters as dense regions in the data space, separated by sparse region [1]. The most important task in this technique is to find the dense regions in dataset based on density. Generally, the density of an object say d can be measured by the number of objects that is close to d . DBSCAN finds the core objects that have dense neighbourhoods. These core objects and their neighbourhoods are then connected to form group of dense regions as clusters.

In order to determine which instances belong together in a cluster, DBSCAN uses the Euclidean distance metrics [9], but, unlike k-means, there is no need to supply the number of desired clusters as it clusters data automatically, find arbitrarily shaped clusters and incorporate a notion of outliers [2]. A cluster is defined as containing at least a minimum number of points, every pair of points which either lies within a user-specified distance (ϵ) connected by a series of points in the cluster that each lie within a distance ϵ of the next point in the chain. Smaller values of ϵ yield denser clusters, this is due to the fact that instances must be closer to one another to belong to the same cluster [2]. Depending on the value of ϵ and the size of cluster that is minimum, it is possible to have some instances that will not belong to any cluster but will only appear scattered around, these are considered as outliers [2].

Although, DBSCAN can cluster objects given input parameters such as ϵ (the maximum radius of a neighbourhood) and $Minpts$ (the minimum number of points required in the neighbourhood of a core object), it expects the user to select parameter values that can lead to discovery of acceptable clusters. Unfortunately, such parameter settings are usually empirically set and difficult to determine, as most algorithms are sensitive to the choice of parameter values. It is more tasking most especially when dealing with real-world, high dimensional dataset.

Input:

V: a data set containing n objects or values

ϵ : the radius parameter

$Minpts$: the neighbourhood density threshold.

Output:

Density-based clusters

Method:

1. mark all objects as unvisited
 2. *do*
 3. select an unvisited object R , randomly;
 4. mark R as visited
 5. if the ϵ -neighbourhood of R has at least $Minpts$ objects
 6. create a new cluster C , and add up R to C ;
 7. let the set of objects be N in the ϵ -neighbourhood of R ;
 8. for each point R^j in N
 9. if R^j is unvisited
 10. mark R^j as visited;
 11. if the ϵ -neighbourhood of R^j has at least $MinPts$ points, add those points to N ;
 12. if R^j still does not belong to any cluster, add R^j to C ;
 13. end for
 14. output C ;
 15. else mark R as noise;
 16. *until* no object is unvisited;
-

Figure 1. DBSCAN algorithm adapted from [1].

3. RELATED WORK

Anomaly detection in dataset is widely reported in the literature. It mainly focuses on determining the patterns in data that do not conform to expected behaviour; most times, the non-conforming patterns are often referred to as outliers, discordant observations, exceptions or contaminants in many application domains [8]. Study in [7] proposed Linfty norm for removal of outliers. The study begins by minimizing the cost function using standard optimization techniques. The main drawback of using Linfty norm minimization as identified in the study is its weak robustness to outliers.

Jiang et al. [10] proposed a two-phase clustering algorithm. The study first modified the traditional k-means algorithm by using heuristics. The results from the heuristics showed that the data points in the same cluster may be most likely all outliers or all non-outliers.

The small clusters in the study and the tree with less number of nodes are selected and regarded as outliers. Also, in the course of mining outliers in database with a view to removing exceptional objects that deviate from the rest of the dataset, Jin et al. [11], proposed ranking of outliers based on a symmetric neighborhood relationship. According to the authors, the proposed technique was to avoid the likely problem of having a wrong estimate when outliers are in the location where the density distributions in the neighbourhood are significantly different. The study reported results analysis that shows the effectiveness of the technique in the ranking of outliers. In a study to isolate outliers in dataset [12], an automated detection of outliers that models the human perception of exceptional values was proposed, using the concept of fuzzy set theory. The paper discussed the procedures for detecting outliers in both discrete and continuous data.

In contrary to low frequency of a value which usually may be classified as outliers when found in a large dataset, the study identify some additional factors from human perceptions of an outlying discrete value. These include the number of records where the value occurs, the number of records with a non-null value, and the total number of frequent occurrence of distinct values within an attribute.

Similarly, the novelty detection reported in the study proposed in [13] has the same concept has outliers detection, as authors normally use the word novelty detection or novelty recognition interchangeably. Augusteijn and Folkert [13] explored the applicability of a probabilistic neural network technique to detect the novelty or patterns. The study is not strictly on outlier detection, but is a classification problem that applied the concept deviation from a specific pattern using the machine learning technique to remotely classify sensed data.

Hawkins et al. [14], proposed replicator neural networks which provide a measure that detects outliers from data records. The study construct model using a sample dataset and the model was used to develop score for outlyingness. The paper identifies cluster labels for each data record which help to interpret the resulting outliers. The present study uses DBSCAN algorithm to detect outliers using the density-based clustering approach. The study has demonstrated the use of the algorithm in a rapidminer software environment.

4. MATERIAL AND METHODS

In this section, the approach used for the detection of outliers is discussed. Dataset from an open repository, UCI [15] is modelled for this study.

4.1 Experimentation

The dataset originally stored in excel file is read into the main process environment of Rapidminer Software. Four operators are used for the model construction as shown in Figure 2. While the import operator read the source file into an excel file, the export operator writes the outputs result into another excel file.

The DBSCAN is sensitive to missing value, therefore, an operator is introduced to replace the missing values using the average of the existing values. The complete sets of data is received by the clustering operator which uses the DBSCAN algorithm to deduce relevant clusters. Unlike some other clustering techniques such as k-means and k-medoids that needs to be told about the number of clusters they should look for, DBSCAN only need to infer the number of clusters that are present, thus, it generate the clusters automatically.

The resulting clusters in this paper apart from been represented graphically as shown in Figure 3, the output is also written to excel file in order to clearly show the noisy and clean data values present in each cluster. The parameter settings used for clustering operator in the setup is shown in Table 1.

Algorithm	Parameters	Settings
DBSCAN	Min points	5
	Epsilon	1.0
	Measure type	Mixed measure
	Mixed measure	MixedEuclidean Distance

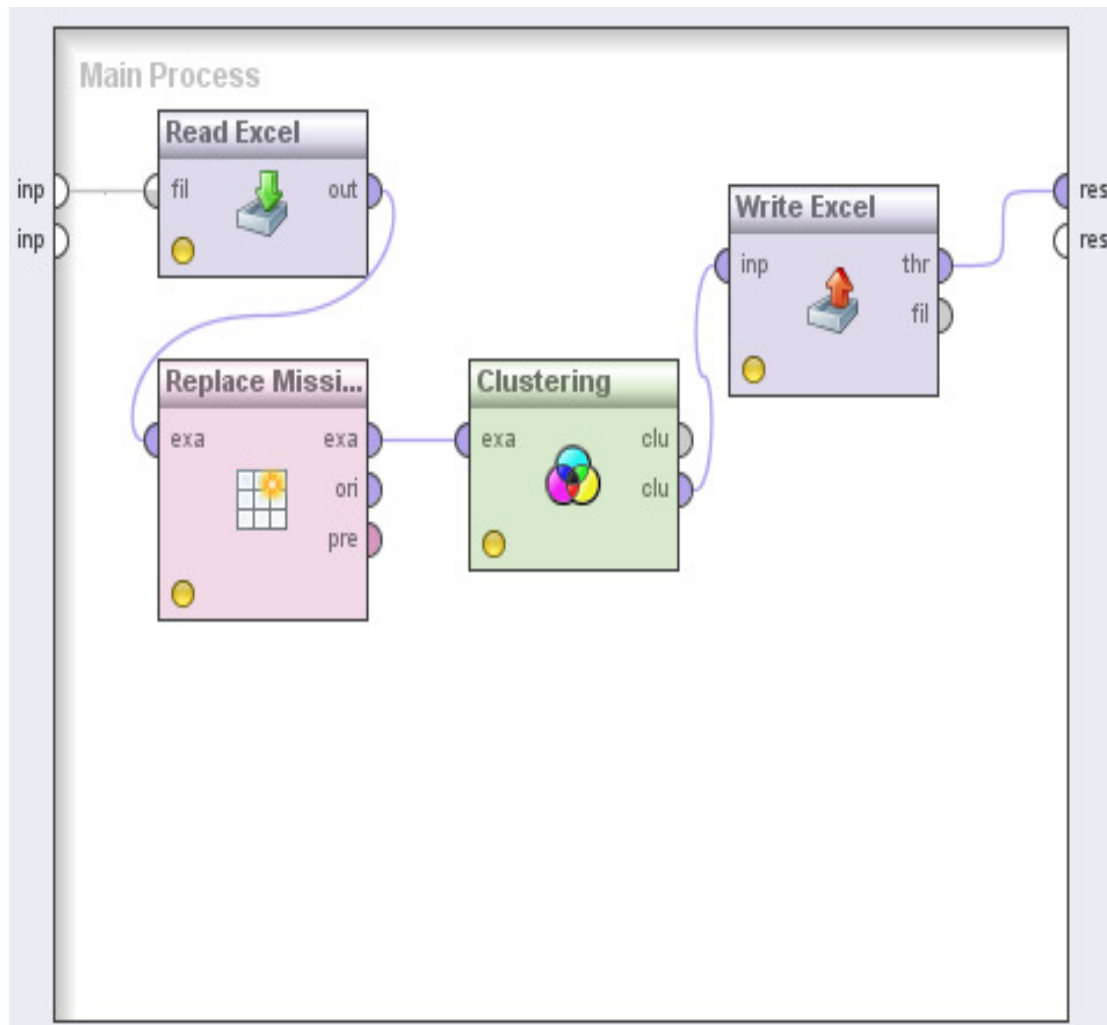
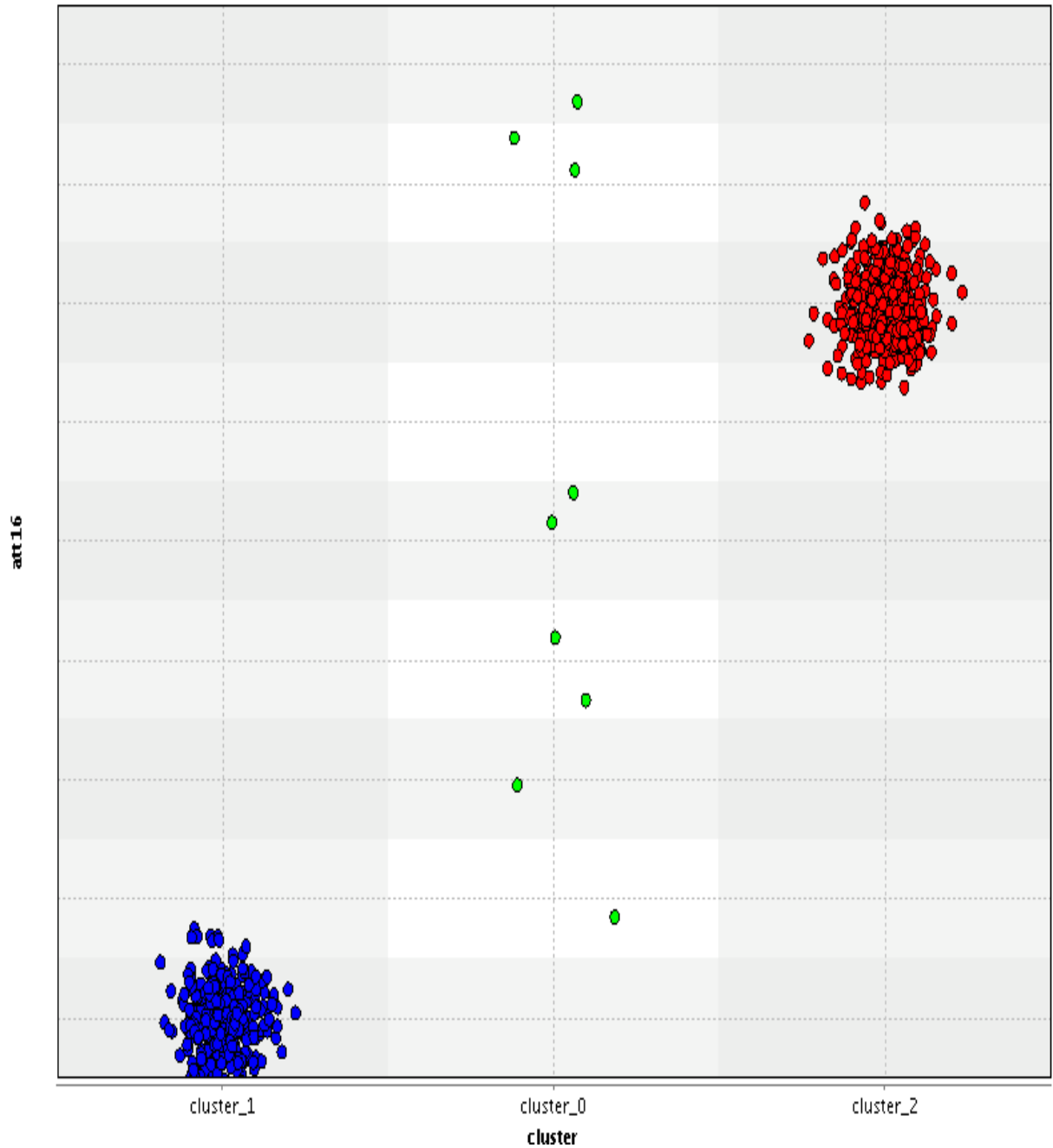


Figure 2. Anomaly detection experimental setup using DBSCAN

For clustering of dataset, DBSCAN essentially needs input parameters such as the maximum radius of a neighbourhood (ϵ) and the minimum number of points that is required in the neighbourhood of a core object (*Minpts*). Table 1 shows the detail parameter settings of DBSCAN operator in Rapidminer.

5. RESULTS AND DISCUSSION

In this section, the output results from the experiment is presented. The algorithm clustered the dataset into three. Clusters 1 and 2 appears very compact, which indicate that each of the cluster has high degree of similarities with itself and shares some features in common. The “cluster 0” is automatically assigned by DBSCAN and it corresponds to anomalies data, the data points appears scattered. These are the points in whose epsilon-neighbourhood that is less than minimum points are found. The normal data are separated from outliers, the noise that is so few and appears scattered as shown in Figure 3.

cluster ● cluster_1 ● cluster_0 ● cluster_2**Figure 3. The clusters generated**

The individual outlier is further represented in a chart (see Figure 4); the content of cluster 1 and 2 forms the desired clean data set.

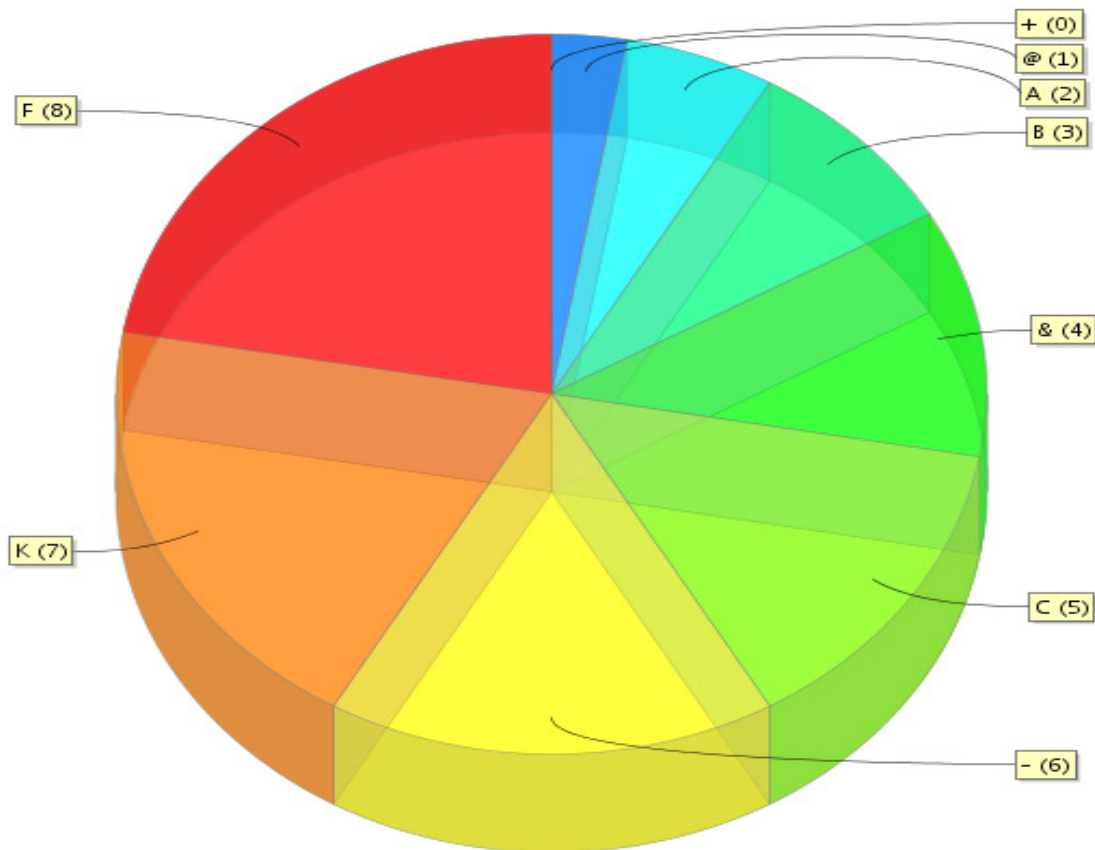


Figure 4. The outliers clustered out from the original dataset.

8. CONCLUSION

In this paper, the anomalies are detected in a dataset using the clustering approach. The DBSCAN algorithm automatically detects data that appear abnormal, while others that shows some closeness in density and free of noise are clustered as normal data. Since there are several other features that need to be considered in the clustering process, normal data points may not necessarily be in the same cluster. The main goal of clustering dataset is to understand the microscopic structure of objects and the actual relationships that exist among the objects. The most promising way to achieve this is through the use of relevant clustering techniques. Most of these techniques can be divided into three categories: those based on distances among objects, those based on density of objects or those based on probability distributions; these are otherwise referred to as distance-based, density-based and distribution-based clustering respectively [5].

Regular geometrical clustering technique such as k-means is sensitive to non-numeric data since it based its clustering strategy on distances among objects, it is therefore, not a usual choice for anomaly detection. The present paper has demonstrated a density-based clustering techniques to unveil anomalies or outliers in datasets. The importance of removing outliers from sets of data is so enormous; one important reason for embarking on such task is to serve as a preventive measure so as not to contaminate the original dataset. The consequence may even be disastrous when noisy dataset is relied upon for model building.

REFERENCES

- [1] J. Han, M. Kamber, and J. Pei, DATA MINING Concepts and Techniques, 3rd ed.: Morgan Kaufmann, Elsevier, USA, 2012.
- [2] H. Witten, E. Frank, and M. A. Hall, DATA MINING Practical Machine Learning Tools and Techniques, 3rd Edition ed.: Morgan Kaufmann, 2011.
- [3] T. Fawcett and F. Provost, "Activity monitoring: Noticing interesting changes in behavior," in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 53-62.
- [4] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," Artificial Intelligence Review, vol. 22, pp. 85-126, 2004.
- [5] D. Skillicorn, Understanding Complex Datasets: Data Mining with Matrix Decompositions. USA: Taylor & Francis Group, 2007.
- [6] M. Negnevitsky, Artificial Intelligence A Guide to Intelligent Systems: Pearson Education Limited, 2011.
- [7] K. Sim and R. Hartley, "Removing Outliers Using the Linfty Norm," in Computer Vision and Pattern Recognition, IEEE Computer Society Conference 2006, pp. 485-494.
- [8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys (CSUR), vol. 41, p. 15, 2009.
- [9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Second international Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, 1996.
- [10] M.-F. Jiang, S.-S. Tseng, and C.-M. Su, "Two-phase clustering process for outliers detection," Pattern recognition letters, vol. 22, pp. 691-700, 2001.
- [11] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in Advances in Knowledge Discovery and Data Mining, ed: Springer, 2006, pp. 577-593.
- [12] M. Last and A. Kandel, "Automated detection of outliers in real-world data," in Proceedings of the second international conference on intelligent technologies, 2001, pp. 292-301.
- [13] M. Augusteijn and B. Folkert, "Neural network classification and novelty detection," International Journal of Remote Sensing, vol. 23, pp. 2891-2902, 2002.
- [14] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in Data Warehousing and Knowledge Discovery, ed: Springer, 2002, pp. 170-180.
- [15] UCI open repository dataset: . Available: <http://archive.ics.uci.edu/ml/datasets.html>