Relocating Local Outliers Produced by K-means and K-medoids Using Local Outlier Rectifier V.2.0

Rogelio O. Badiang Jr.
Technological Institute of the Philippines
938 Aurora Blvd., Cubao
Quezon City, Philippines
Email: rj.badiang@gmail.com

Bobby D. Gerardo
West Visayas State University
Luna St., La Paz
Iloilo City, Philippines
Email: bgerardo@wvsu.edu.ph

Ruji P. Medina
Technological Institute of the Philippines
938 Aurora Blvd., Cubao
Quezon City, Philippines
Email: ruji.medina@tip.edu.ph

***Abstract*—The extensive growth in the field of information and communication technology allows easy capture of massive amounts of valuable data in different areas. These data are used in various data mining techniques. However, in some cases, the presence of outliers in the dataset exists. One of the categories of an outlier is the local outlier. Local outliers are data points that deviate locally from the cluster center. They occur when the cluster center, known as centroid or medoid, cannot represent all the data members in the cluster. The unrepresented data are mistakenly classified to their closest clusters, making them local outliers. With this, the study aims to address the problem of local outliers produced by K-means and K-medoids. The Local Outlier Rectifier V.2.0 (LOR V.2.0) is a method used to relocate local outliers to their correct clusters. The simulations show that when LOR V.2.0 is partnered with K-means, it was able to relocate 35.37%, 34.78%, 25%, and 12.28% local outliers of Ionosphere, Breast Cancer Wisconsin, Iris, and Breast Cancer Coimbra datasets, respectively. On the contrary, when LOR V.2.0 is partnered with K-medoids, 29.67% of Breast Cancer Wisconsin, 29.11% of Ionosphere, 25.0% of Iris, and 10.34% of Breast Cancer Coimbra local outliers were transferred to their correct clusters. The result also indicates that the method works better when partnered with K-means.**

Keywords—local outlier; mahalanobis distance; median absolute deviation; k-means; k-medoids.

I. Introduction

With the massive development in the area of information and communication technology, massive amounts of valuable data being captured are also expanding in different domains. Concurrently, these data are used in various data mining techniques such as data clustering. However, the presence of some rare and extraordinary data that deviates from other surrounding data, known as outliers [1], [2], occurs.

Generally, outliers are categorized into three types, specifically global, contextual, and collective outliers [3]. Global outliers are data that deviate from the entire dataset [4], whereas a collective outlier refers to the subset of observations that significantly deviates from the whole dataset [5]. On the other hand, contextual outliers, also known as conditional outliers, are a generalization of local outliers [3]. Local outliers are data that deviates locally from the cluster center [6].

In partition methods, cluster center represents each data point in every cluster [7]. The outliers occur when data points have huge dissimilarity to each other, and the cluster center cannot represent all of them. The unrepresented data points will be mistakenly clustered to their nearest clusters.

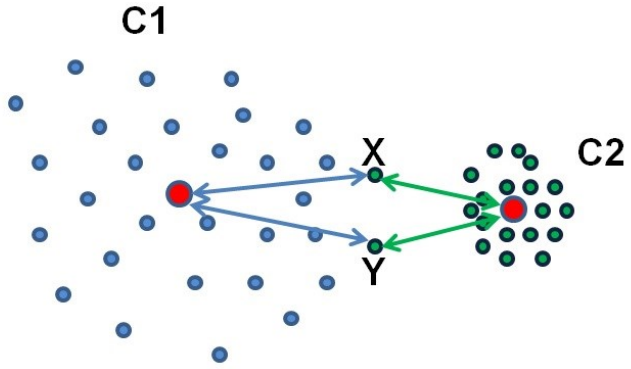

Fig. 1. Dataset Overview

Shown in Fig. 1 is a dataset clustered into two, namely C1 and C2. Supposedly data points X and Y belong to C1 however, clustered to C2 because the data points are nearer to the cluster center of C2, making both the local outliers.

This study introduces the Local Outlier Rectifier (LOR) V.2.0 method that will relocate local outliers to their correct clusters. Furthermore, this study also compares the clustering performance of K-means and K-medoids with and without LOR V.2.0 as their partner.

Likewise, this paper is breakdown into five sections. Section 2 tackles about literature review, and section 3 explains the proposed method. The results of the simulation are discussed in section 4. Finally, presented in section 5 is the conclusion of the study.

II. Literature Review

A. Outlier Detection Approaches

In general terms, global outliers are called outliers. These are data points that do not belong to any cluster in the dataset. On the other hand, local outliers are data points of one cluster that are members of other clusters [8]. Additionally, it is defined as data points that deviate from its local neighborhood concerning the densities of the neighborhoods [6].

There are several approaches used in identifying outliers, and one of which is the distance-based approach. The approach made use of the distance function in relating each data points in the dataset [9]. The data points far from other neighboring data points are assumed as outliers. Mahalanobis distance, Euclidean distance, and other measures of dissimilarity are the appropriate distance measures for this approach [10]. There are several outlier detection algorithms based on the distanced-based approach being developed. Knorr and Ng introduced the first distance-based outlier detection technique in 1998 [11].

Moreover, local outliers are identified using LDOF or Local Distanced-based Outlier Factor that measures how much a data point differs from its neighborhood [12]. The data point having a higher LDOF indicates that it is more likely to be an outlier. Because of the computational inefficiency, an improvement of LDOF known as Pruning-based LDOF is proposed [9]. The concept of this algorithm is to prune half of the clusters having less clusterldof value. The LDOF of the other half is calculated, which is the basis for identifying outliers.

Alternatively, density-based approach gained popularity over the years. The approach addresses the problems of varying levels of cluster density in the dataset encountered by distance-based approach. The approach's basic idea is that the neighborhood's density of one data point is correlated with that of the neighborhood of its neighbor. The data point is assumed as an outlier if the difference between densities is significant [10]. The implementation of this approach is seen in the development of several algorithms. Some of these are outlier detection method based on multi-dimensional clustering and local density (ODBMCLD) [13], relative density-based outlier score (RDOS) [1], density-based outlier detection technique [14], and density-based local outlier detection on uncertain data [15].

Another outlier detection approach is using distribution or statistical-based. In the statistical approach, it is assumed that the dataset follows a normal distribution and any data point that deviates from such distribution is an outlier [16]. If the dataset follows a normal distribution, an outlier is considered when a data point is more than three standard deviations [17] from the mean. On the contrary, Wilcox mentioned that when a value is more than two standard deviations from the mean, then it is an outlier [18].

B. Partitioning Methods

Partition method is one of the categories of clustering algorithms. It divides data objects into multiple initial clusters and constantly relocates data objects to their closest cluster or centroid using the dissimilarity criterion [19]. The result of the clustering satisfies two conditions, every data point is a member of one cluster only, and each cluster has a minimum of one data member [20], [21], [22].

K-means and K-medoids are two of the classifications of partition methods. K-means clustering updates the cluster center by iterative computation until the convergence criteria are met [23]. The cluster center is the mean of all members of each cluster [20].

On the contrary, K-medoids clustering is an improvement of k-means that utilizes medoids instead of the mean in representing each cluster. The medoids are the actual data points in the dataset [24]. Partition Around Medoid (PAM) is one of the representatives of K-medoids [25].

C. Mahalanobis Distance

In multivariate statistics, one of the measures is Mahalanobis distance. Shown in (1) is the Mahalanobis distance formula, where o is the object from the dataset, o is the mean vector, and S is the covariance matrix [3].

$$MDist(o, \bar{o}) = \sqrt{(o - \bar{o})T\ S\ -1\ (o - \bar{o})}\ , \quad (1)$$

The distance can be a basis in determining whether a data point is an outlier or whether a data point is a member of a cluster or not [26].

D. Median Absolute Deviation

Many outlier detection methods rely on distribution properties like mean and variance. However, the presence of outliers has an impact on the shift of the mean and variance resulting in the"true" outliers undetected [27]. With this, the robust statistics aim to design statistics that are more resistant to outliers. One of which is the Median Absolute Deviation or MAD that estimates the dispersion of the data in a dataset [28].

MAD only involves calculating the median of absolute deviation from the median given the equation below [29].

$$\mathrm{MAD_n} = \mathrm{b\ med_i}\ |\ \mathrm{X_i} - \mathrm{med_j X_j}\ |\ , \quad (2)$$

where med is the sample median. The constant b in (2) is set to 1.4826 in the case of the usual parameter θ at Gaussian distributions [30].

E. Clustering Accuracy : Performance Measure

One of the techniques used in measuring the clustering performance is accuracy. It refers to the percentage of the data points in clustering results that were correctly recovered. The clustering accuracy is expressed as

$$r = 100\frac{\sum_{i=1}^{k} a_i}{N}\ , \quad (3)$$

where a_i is the number of data points correctly classified in every cluster, and N is the number of data points in the entire dataset [31].

III. Proposed Method: Local Outlier Rectifier V.2.0

The local outliers produced by K-means and K-medoids are relocated using the proposed method Local Outlier Rectifier V.2.0 or LOR V.2.0. The local outliers are the data points incorrectly clustered by the above algorithms. The incorrect clustering happened when the cluster center can not represent data points far from it. The nearest neighboring cluster center will represent the data points, making them members of that cluster [7].

Shown in Fig. 2 are the major processes of the LOR V.2.0.

The proposed method takes the clustering results of K-means or K-medoids as input. On each cluster input, the potential local outliers are identified. The process of identification requires the following steps:

1. Get the Mahalanobis distance of each data point from their median and calculate the Median Absolute Deviation.
2. Identify potential local outliers.
3. Remove the potential local outliers from their clusters.

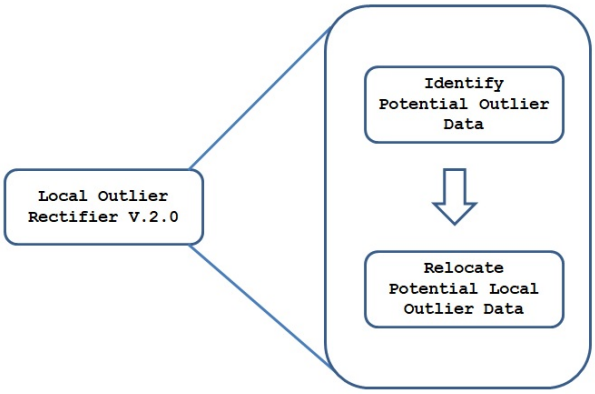

Fig. 2. LOR V.2.0 Major Processes

The term potential is being used because not all identified data points are real outliers. The basis for the identification of potential outliers is those data points having Mahalanobis distance which is not in the range of Median ± Median Absolute Deviation of the Mahalanobis distances.

The Median ± Median Absolute Deviation is based on the idea that the Median Absolute Deviation approximately represents the Standard Deviation of the normal distribution [32], [33]. Likewise, the one standard deviation, Mean ± Standard Deviation, threshold in identifying outliers is used on studies in identifying discrete faults in synthetic data [34] and identifying malaria epidemic in Africa [35].

The last major process is relocating the potential outliers to their appropriate clusters. The process requires the following steps:
On every cleaned cluster,

1. Merge the potential local outliers one at a time.
2. Get the Mahalanobis distance of each data point from their median.
3. Relocate the local outlier data.

Those potential outliers that are true outliers will be relocated to where they rightfully belong while those who are not will remain to their original clusters. They will be relocated by checking their Mahalanobis distance from each cluster. The cluster having the datum's smallest distance from their median will be its new cluster.

IV. Simulation Results

The simulation uses datasets from kaggle.com and Machine Learning Repository (archive.ics.uci.edu). The Breast Cancer Coimbra, Breast Cancer Prediction (Wisconsin), Ionosphere, and Iris datasets are the datasets being used. The general information on the four datasets is shown in Table I.

All datasets were clustered using PAM and K-means. The results of the algorithms were fed to the LOR V.2.0 method as inputs. The tools used during the simulations are the R programming language and RStudio as an IDE.

TABLE I. DATA SET INFORMATION

Dataset	Samples	Features	Classes
Breast Cancer Coimbra	116	9	2
Breast Cancer Wisconsin	569	5	2
Ionosphere	253	11	2
Iris	150	4	3

A. K-means and LOR V.2.0

Since the clustering result of K-means changes every time it is run, the algorithm is executed ten times. The average accuracy, the number of local outliers, and the number of local outliers correctly relocated are computed.

Table II shows the accuracy of K-means and the number of local outlier data. Moreover, Table III shows the new accuracy after taking the clusters produced by K-means as input during the execution of the LOR V.2.0 method.

TABLE II. THE AVERAGE K-MEANS CLUSTERING RESULTS FOR 10 EXECUTIONS

Dataset	Accuracy	No. of Local Outliers
Breast Cancer Coimbra	50.86 %	57
Breast Cancer Wisconsin	83.83 %	92
Ionosphere	67.59 %	82
Iris	89.33 %	16

TABLE III. THE AVERAGE K-MEANS + LOR V.2.0 CLUSTERING RESULTS FOR 10 EXECUTIONS

Dataset	Accuracy	No. of Correctly Relocated Local Outliers
Breast Cancer Coimbra	56.90%	7
Breast Cancer Wisconsin	89.46%	32
Ionosphere	78.94%	29
Iris	92.00%	4

Notably, the accuracy of all the datasets increased. The Ionosphere dataset got the highest gain of 11.35 percentage points while Iris dataset got the smallest increase of only 2.67 percentage points. The Breast Cancer Coimbra and Breast Cancer Wisconsin have a percentage points increase of 6.03 and 5.62 accordingly.

In addition, there is 35.37 % of Ionosphere, 34.78 % of Breast Cancer Wisconsin, 12.28 % of Breast Cancer Coimbra, and 25.0 % of Iris local outliers correctly relocated to their appropriate clusters.

B. PAM and LOR V.2.0

Table IV illustrates the accuracy of the PAM clustering

algorithm and the number of local outliers wrongly classified to other clusters. Table V also illustrates the accuracy of clustering performance after the execution of the LOR method.

TABLE IV. PAM CLUSTERING RESULTS

Dataset	Accuracy	No. of Local Outliers
Breast Cancer Coimbra	50.0 %	58
Breast Cancer Wisconsin	84.01 %	91
Ionosphere	68.77 %	79
Iris	89.33 %	16

TABLE V. PAM + LOR V.2.0 CLUSTERING RESULTS

Dataset	Accuracy	No. of Correctly Relocated Local Outliers
Breast Cancer Coimbra	55.17 %	6
Breast Cancer Wisconsin	88.75 %	27
Ionosphere	77.87 %	23
Iris	92.00 %	4

Similar to the prior findings, there is a noticeable increase in accuracy after the execution of the LOR V.2.0 method. Still, Ionosphere dataset got the highest increase in accuracy with 9.09 percentage points. On the other hand, Iris dataset with the smallest increase only got 2.67 percentage points. Breast Cancer Coimbra got 5.17 percentage points gain while Breast Cancer Wisconsin got 4.75 percentage points.

Additionally, 29.67% or 22 out of 91 local outliers from Breast Cancer Wisconsin dataset were transferred to their appropriate clusters. The Ionosphere, Breast Cancer Coimbra, and Iris datasets got 29.11%, 10.34%, and 25.0% of the local outliers were correctly transferred, respectively.

Likewise, the LOR V.2.0 method achieved better accuracy when partnered with K-means. The simulation shows that the partners outperformed the PAM + LOR V.2.0 on three datasets except for Iris. The accuracies on Iris dataset are equal in both partners.

C. Comparison of LOR V.2.0 to Other Algorithms

There are several modifications of K-means and K-medoids that aim to improve the clustering performance, specifically the accuracy. The clustering accuracy of LOR V.2.0 on Iris dataset, shown in Table VI, and Ionosphere dataset, shown in Table VII, are compared to other algorithms since they got the smallest and the highest percentage increase, respectively.

The LOR V.2.0 does not perform least among all the algorithms. It is observed that LDA-Km algorithm got the highest accuracy with 98.0% on Iris dataset. It has six percentage points difference to LOR V.2.0 partnered to both K-means and PAM.

In contrast, K-means + LOR V.2.0 is 7.74 percentage points higher than LDA-Km on Ionosphere dataset. Also, a similar observation in the case of PAM + LOR V.2.0, which is 6.64 percentage points higher.

Additionally, the algorithm with the highest accuracy on the Ionosphere dataset is Accelerated K-means Clustering with 93.4%. It is 14.46 percentage points higher than K-means + LOR V.2.0. It is also 15.53 percentage points higher than PAM + LOR V.2.0. Despite the huge difference in Ionosphere dataset, the two LORs have higher accuracy than Accelerated K-means Clustering on Iris dataset with both 1.4 percentage points difference.

TABLE VI. CLUSTERING ACCURACY OF DIFFERENT ALGORITHMS ON IRIS DATASET

Algorithm	Accuracy
Improved K-means [36]	90.6 %
IKCBD [37]	92.67 %
Accelerated K-means Clustering [38]	90.6 %
K-means Clustering Algorithm - Based on Improved PSO [39]	90.13 %
Simple and Fast Algorithm for K-medoids [40]	92.0 %
LDA-Km [41]	98.0 %
K-means + LOR V.2.0	92.0 %
PAM + LOR V.2.0	92.0 %

TABLE VII. CLUSTERING ACCURACY OF DIFFERENT ALGORITHMS ON IONOSPHERE DATASET

Algorithm	Accuracy
Accelerated K-means Clustering [38]	93.4%
LDA-Km [41]	71.2%
K-Means(fast) [42]	71.23%
K-Means(kernel) [42]	55.56%
K-Means(kernel) Optim. [42]	64.10%
K-Medoids Optim. [42]	72.36%
K-means + LOR V.2.0	78.94%
PAM + LOR V.2.0	77.87%

V. CONCLUSION

In this paper, the problem of local outliers produced by K-means and K-medoids are addressed. The problem of local outliers arises when the cluster centers cannot represent all its members. It was found out in the simulation that the proposed LOR V.2.0 method successfully relocated the local outliers to its right clusters. Thus, the accuracy of clustering increases. Furthermore, the LOR V.2.0, when partnered with K-means, performed better than PAM as its partner.

ACKNOWLEDGMENT

Above all, we are thanking our Lord God Almighty for giving us knowledge and wisdom, good health, and the opportunity to carry through this research study.

We are also extending our sincere and warm thanks to all the people who helped us with this journey, especially to all the TIP professors, panelists, and friends.

This acknowledgment will not be completed without

giving a special thanks to our family members. We are grateful for the love and support that you have given us in all our endeavors.

REFERENCES

[1] B. Tang and H. He, "A local density-based approach for outlier detection", Neurocomputing, vol. 241, pp. 171-180, 2017. doi: 10.1016/j.neucom.2017.02.039

[2] Z. Gao, "Application of cluster-based local outlier Factor Algorithm in Anti-Money Laundering", 2009 International Conference on Management and Service Science, 2009. doi: 10.1109/icmss.2009.5302396

[3] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Elsevier, 2012.

[4] S. Chawla and A. Gionis, "k-means: a unified approach to clustering and outlier detection", Proceedings of the 2013 SIAM International Conference on Data Mining, 2013. doi: 10.1137/1.9781611972832.21.

[5] P. Skrabanek and N. Martinkova, "Extraction of outliers from imbalanced sets", Hybrid Artificial Intelligent Systems, 2017. doi: 10.1007/978-3-319-59650-1

[6] M. Breunig, H. Kriegel, R. Ng and J. Sander, "LOF: identifying density-based local outliers.", ACM SIGMOD Record, vol. 29, no. 2, pp. 93-104, 2000. doi: 10.1145/335191.335388

[7] D. Sisodia,L. Singh, S. Sisodia, and K. Saxena,"Clustering techniques: a brief survey of different clustering algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET),pp.82-87,2012.

[8] Y. Zhou, H. Yu and X. Cai, "A novel k-Means algorithm for clustering and outlier detection", 2009 Second International Conference on Future Information Technology and Management Engineering (FITME 2009), Sanya, 2009. doi: 10.1109/FITME.2009.125

[9] R. Pamula, J. Deka and S. Nandi, "Distance-based fast outlier detection method", 2010 Annual IEEE India Conference (INDICON), 2010. doi: 10.1109/indcon.2010.5712706

[10] S. Rakhe,and A. Vaidya, "A Survey on different unsupervised techniques to detect outliers." International Research Journal of Engineering and Technology (IRJET) Volume 2 (2015).

[11] E. Knorr, and R. Ng, "Algorithms for mining distance-based outliers in large datasets." Proceedings of the international conference on very large data bases, pp. 392-403. Citeseer, 1998.

[12] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data.", Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 813-822. Springer, Berlin, Heidelberg, 2009.

[13] Z. Shou, M. Li and S. Li, "Outlier detection based on multi-dimensional clustering and local density", Journal of Central South University, vol. 24, no. 6, pp. 1299-1306, 2017. doi: 10.1007/s11771-017-3535-4

[14] R. Gupta and K. Pandey, "Density-based outlier detection technique", Advances in Intelligent Systems and Computing, pp. 51-58, 2016. doi: 10.1007/978-81-322-2755-7_6

[15] K. Cao, L. Shi, G. Wang, D. Han and M. Bai, "Density-based Local outlier detection on uncertain data", Web-Age Information Management, pp. 67-71, 2014. doi: 10.1007/978-3-319-08010-9_9

[16] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, Statistical outlier detection using direct density ratio estimation, Knowl. Inform. Syst., vol. 26, no. 2, pp. 309336, 2011.

[17] K. Rosen, "Handbook of discrete and combinatorial mathematics", 2nd ed., CRC Press, 2018.

[18] R. Wilcox, "Fundamentals of modern statistical methods" ,New York: Springer, 2010.

[19] S. Baadel, F. Thabtah, and J. Lu, 2016. "Overlapping clustering: a review", 2016 SAI Computing Conference (SAI), 2016. doi: http://dx.doi.org/10.1109/sai.2016.7555988

[20] A. Fahad et al., "A Survey of clustering algorithms for big data: taxonomy and empirical analysis", IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, 2014. doi:10.1109/tetc.2014.2330519

[21] M. Shah and S. Nair, "A survey of data mining clustering algorithms", International Journal of Computer Applications, vol. 128, no. 1, pp. 1-5, 2015. doi: 10.5120/ijca2015906404

[22] J. Swarndeep Saket, and S. Pandya. "An overview of partitioning algorithms in clustering techniques." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no.6, 2016.

[23] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms", Annals of Data Science, vol. 2, no. 2, pp. 165-193, 2015. doi: 10.1007/s40745-015-0040-1

[24] A. Bhat, "K-medoids clustering using partitioning around medoids for performing face recognition", International Journal of Soft Computing, Mathematics and Control, vol. 3, no. 3, pp. 1-12, 2014. doi:10.14810/ijscmc.2014.3301

[25] X. Jin and J. Han, "K-medoids clustering",Encyclopedia of Machine Learning and Data Mining, 2019. doi:http://dx.doi.org/10.1007/9781-4899-7687-1_432

[26] R.Brereton, "The mahalanobis distance and its relationship to principal component scores", Journal of Chemometrics, 29(3),pp.143-145, 2015.

[27] S. Sabade and D. Walker, "Evaluation of effectiveness of median of absolute deviations outlier rejection-based I/sub DDQ/ testing for burn-in reduction", Proceedings 20th IEEE VLSI Test Symposium (VTS 2002). doi: 10.1109/vts.2002.1011115

[28] P. Kersten, "Fuzzy order statistics and their application to fuzzy clustering", IEEE Transactions on Fuzzy Systems, vol. 7, no. 6, pp. 708-712, 1999. doi: 10.1109/91.811239

[29] C. Leys, C. Ley, O. Klein, P. Bernard and L. Licata, "Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median", 2019. doi: http://dx.doi.org/10.1016/j.jesp.2013.03.013

[30] P. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation", Journal of the American Statistical Association, vol. 88, no. 424, pp. 1273-1283, 1993. doi: 10.1080/01621459.1993.10476408

[31] J. Huang, M. Ng, H. Rong and Z. Li, "Automated variable weighting in k-means type clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657-668, 2005. doi:10.1109/tpami.2005.95

[32] A. Comport, M. Pressigout, E. Marchand and F. Chaumette,"A visual servoing control law that is robust to image outliers", Proceedings 2003 IEEERSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453), 2003. doi:10.1109/iros.2003.1250677

[33] N. Chung et al., "Median Absolute Deviation to Improve Hit Selection for Genome-Scale RNAi Screens", Journal of Biomolecular Screening, vol. 13, no. 2, pp. 149-158, 2008. doi: 10.1177/1087057107312035.

[34] S. Das, B. Matthews, A. Srivastava, and N. Oza,"Multiple kernel learning for heterogeneous anomaly detection", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 10, 2010.

[35] T. Abeku, et al., "Malaria epidemic early warning and detection in African highlands", Trends in Parasitology, 20(9), pp.400-405, 2004.

[36] U. Raval and C. Jani, "Implementing & improvisation of K-means clustering algorithm", International Journal of Computer Science and Mobile Computing, vol. 5, no. 5, pp.191-203, 2016.

[37] W. Shunye, "An improved k-means clustering algorithm based on dissimilarity", Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), 2013. doi:10.1109/mec.2013.6885476

[38] P. Jain and B. Buksh, "Accelerated K-means Clustering Algorithm", International Journal of Information Technology and Computer Science, vol. 8, no. 10, pp. 39-46, 2016. doi: 10.5815/ijitcs.2016.10.05

[39] L. Tan, "A Clustering K-means Algorithm Based on Improved PSO Algorithm", 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015. doi: 10.1109/csnt.2015.223.

[40] H. Park and C. Jun, "A simple and fast algorithm for K-medoids clustering", Expert Systems with Applications, vol. 36, no. 2, pp. 3336-3341, 2009. doi: 10.1016/j.eswa.2008.01.039

[41] C. Ding and T. Li, "Adaptive dimension reduction using discriminant analysis andK-means clustering", Proceedings of the 24th international conference on Machine learning - ICML '07, 2007. doi:10.1145/1273496.1273562

[42] I. Rozhnov, V. Orlov, and L. Kazakovtsev, "Ensembles of Clustering Algorithms for Problem of Detection of Homogeneous Production Batches of Semiconductor Devices", CEUR Workshop Proceeding, vol. 2098, pp.338-48, 2018.